Noname manuscript No.
(will be inserted by the editor)

Learning Generalizable Mixed-Precision Quantization via Attribution Imitation

Ziwei Wang · Han Xiao · Jie Zhou · Jiwen Lu

Received: date / Accepted: date

Abstract In this paper, we propose a generalizable mixed-precision quantization (GMPQ) method for efficient inference. Conventional methods require the consistency of datasets for bitwidth search and model deployment to guarantee the policy optimality, leading to heavy search cost on challenging large-scale datasets in realistic applications. On the contrary, our GMPQ searches the mixed-quantization policy that can be generalized to large-scale datasets with only a small amount of data, so that the search cost is significantly reduced without performance degradation. Specifically, we observe that locating network attribution correctly is general ability for accurate visual analysis across different data distribution. Therefore, despite of pursuing higher accuracy and lower model complexity, we preserve attribution rank consistency between the quantized models and their full-precision counterparts via capacity-aware attribution imitation (CAI) for generalizable mixed-precision quantization strategy search, where the capacity of quantized networks is considered to fully utilize the network capacity without insufficiency. Since slight noise in attribution is amplified by discrete ranking operations with significant rank errors, mimicking the attribution ranks of the full-precision models obstructs the quantized networks to correctly locate the attribution. To address this, we further present a robust generalizable mixed-precision quantization (R-GMPQ) method to smooth the attribution for rank error alleviation by hierarchical attribution partitioning, which efficiently partitions the attribution pixels in high spatial resolution and assigns the same attribution value for pixels within a group. Moreover, we propose dynamic capacity-aware attribution imitation (DCAI) to adjust the concentration degree of the attribution according to sample hardness, so that sufficient model capacity is achieved with full utilization for each image. Extensive experiments on image classification and object detection show that our GMPQ and R-GMPQ obtain competitive accuracy-complexity trade-offs with significantly reduced search cost compared to the state-of-the-art mixed-precision networks.

Keywords Mixed-precision quantization · Generalizable compression policy · Attribution rank preservation · Attribution imitation · Hierarchical attribution partitioning

Ziwei Wang[1]
E-mail: wang-zw18@mails.tsinghua.edu.cn
Han Xiao[1]
E-mail: h-xiao20@mails.tsinghua.edu.cn
Jie Zhou[1]
E-mail: jzhou@tsinghua.edu.cn
Jiwen Lu[1,✉]
E-mail: lujiwen@tsinghua.edu.cn

[1] State Key Lab of Intelligent Technologies and Systems, Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, 100084, China

1 Introduction

Deep neural networks have achieved the state-of-the-art performance across a large number of vision tasks such as image classification (He et al., 2016; Huang et al., 2017; Simonyan and Zisserman, 2014), object detection (He et al., 2017a; Liu et al., 2016; Ren et al., 2015), face recognition (Deng et al., 2019; Liu et al., 2017; Wang et al., 2018) and many others. However, the mobile devices with limited storage and computational resources are not capable of processing deep models due to the extremely high complexity. Therefore, it is desirable to

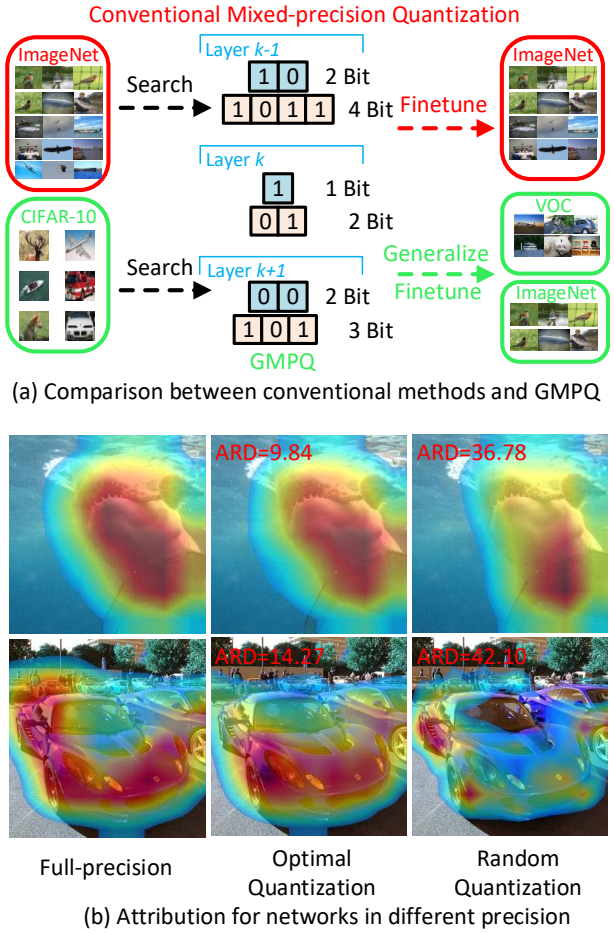


Fig. 1 (a) The difference between conventional mixed-precision networks and our GMPQ. Conventional methods require the consistency of datasets for bitwidth search and model deployment, while our GMPQ searches the optimal quantization policy on small datasets and generalizes it to large-scale datasets. Therefore, the search cost is significantly reduced in our GMPQ. (b) The attribution computed by Grad-cam for full-precision, randomly and optimally quantized networks for example images from ImageNet (top row) and PASCAL VOC (bottom row). Different from random quantization, the optimal quantization policy keeps the similar attribution rank with the full-precision counterparts regardless of the datasets.

design network compression strategy according to the hardware configurations.

Recently, several network compression techniques have been proposed including pruning (He et al., 2017b; Lin et al., 2017; Molchanov et al., 2019), quantization (Liu et al., 2018; Wang et al., 2020a; Zhao et al., 2019), efficient architecture design (Howard et al., 2017; Iandola et al., 2016; Qin et al., 2019) and low-rank decomposition (Denton et al., 2014; Li et al., 2020a; Yu et al., 2017). Among these approaches, quantization constrains the network weights and activations in limited bitwidth for memory saving and fast processing. In order to fully utilize the hardware resources, mixed-precision quantization (Cai and Vasconcelos, 2020; Dong et al., 2019c; Wang et al., 2019a) is presented to search the bitwidth in each layer so that the optimal accuracy-complexity trade-off is obtained. However, conventional mixed-precision quantization requires the consistency of datasets for bitwidth search and network deployment to guarantee policy optimality, which causes significant search burden for automated model compression on large-scale datasets such as ImageNet (Deng et al., 2009). For example, it usually takes several GPU days to acquire the expected quantization strategy for ResNet18 on ImageNet (Cai and Vasconcelos, 2020; Wang et al., 2019a).

In this paper, we present a GMPQ method to learn generalizable mixed-precision quantization strategy via attribution rank preservation for efficient inference. Unlike existing methods which require the dataset consistency between quantization policy search and model deployment, our method enables the acquired quantization strategy to be generalizable across various datasets. The quantization policy searched on small datasets achieves promising performance on largescale datasets, so that policy search cost is significantly reduced. Figure 1(a) shows the difference between our GMPQ and conventional mixed-precision networks. More specifically, we observe that correctly locating the network attribution benefits visual analysis for various input data distribution. Attribution has been proven to be effective in computer vision and machine learning with transferability requirements (Gao et al., 2023; Wang et al., 2021; Zunino et al., 2021), where feature attribution from different well-performed networks is very similar for the same input instance regardless of the data distribution. Inspired by this, we also observe that the feature attribution acquired in full-precision networks and quantized ones with the optimal policy is very similar, as depicted in Figure 1(b) and Table 1 respectively. Figure 1(b) demonstrates the attribution computed by Grad-cam (Selvaraju et al., 2017) for mixed-precision networks with optimal and random quantization policy and their full-precision counterparts, where the optimal quantization policy is acquired by the mixed-precision quantization search method EdMIPS (Cai and Vasconcelos, 2020) performed on corresponding datasets. Table 1 demonstrates the attribution rank difference (ARD) between attribution in the quantized and full-precision networks. The mixed-precision networks with the optimal bitwidth assignment acquire more consistent attribution rank with the full-precision model regardless of data distribution, which also indicates high generalization ability across different distribution. Therefore, despite of considering model accuracy and complexity,

Table 1 The average attribution rank difference across all test images for different quantization policies.

	ImageNet	VOC	COCO
Optimal	10.03	15.78	28.99
Random	39.94	40.07	78.95

we enforce the quantized networks to imitate the attribution of the full-precision counterparts. Instead of directly minimizing the Euclidean distance between attribution of quantized and full-precision models, we preserve their attribution rank consistency so that the attribution of quantized networks can adaptively adjust the distribution without model capacity insufficiency. Moreover, we also present capacity-aware attribution imitation (CAI) for efficient optimization of attribution rank consistency preservation, where we enforce the attribution in quantized networks to mimic that in full-precision counterparts with adaptive concentration degree decided by the network capacity.

In fact, the attribution ranks in full-precision networks usually contain errors because slight attribution noise can be significantly amplified by ranking operations, which fails to reveal the true region importance. Figure 2(a) shows an example of rank errors in attribution ranks. Since GMPQ enforces the quantized networks to mimic the attribution rank of full-precision counterparts, the rank errors in ranks hinder the quantized networks to correctly locate the attribution. In order to address these limitations, we further propose a robust generalizable mixed-precision quantization (R-GMPQ) method to smooth the attribution for noise alleviation by hierarchical attribution partitioning. More specifically, we hierarchically partition attribution according to semantic similarity of pixels in feature maps, where the pixel groups are decided in different spatial resolution. Following that, we smooth the attribution by the statistics of different partitions. Figure 1(b) demonstrates the difference between GMPQ and R-GMPQ. Meanwhile, the capacity-aware attribution imitation (CAI) assigns the same concentration degree of attribution for samples in various hardness, which fails to fully utilize the model capacity for easy images and causes capacity insufficiency for hard samples. We present dynamic capacity-aware attribution imitation (DCAI) to adjust the attribution distribution adaptively according to sample hardness that is evaluated by the task risk changes during model quantization, so that sufficient model capacity is acquired with full utilization for each input. Compared with the state-of-the-art mixed-precision quantization methods, extensive experiments show that our GMPQ and R-GMPQ obtain competitive accuracy-complexity trade-off with significantly reduced search cost on ImageNet (Deng et al., 2009) for image classification and on PASCAL VOC (Everingham et al., 2010) and COCO (Lin et al., 2014) for object detection.

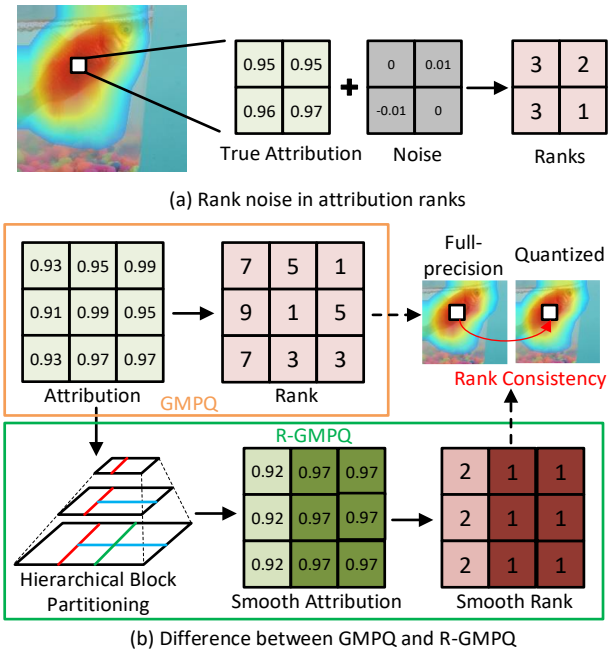


Fig. 2 (a) The attribution rank errors. The attribution noise is negligible in the selected area, which is significantly amplified by ranking operations. (b) The difference between GMPQ and R-GMPQ. GMPQ directly enforces the quantized neural networks to preserve the attribution rank consistency with full-precision counterparts, so that the quantized model fails to locate the attribution correctly due to the attribution rank errors in full-precision networks. R-GMPQ hierarchically partitions the attribution, and smooths the attribution with partition statistics for rank error alleviation.

This paper is an extended version of our conference paper, we make the following new contributions:

1. We propose a new R-GMPQ method based on GMPQ by smoothing the attribution with hierarchical attribution partitioning, so that the attribution rank errors are alleviated for generalizable mixed-precision quantization policy search.
2. We present dynamic capacity-aware attribution imitation to select the optimal attribution concentration degree adaptively according to sample hardness, so that sufficient model capacity is acquired with full utilization for each image.
3. We conducted extensive experiments on image classification and object detection, and the results demonstrate the effectiveness and efficiency of the presented methods.

2 Related Work

In this section, we briefly review three related topics: 1) fixed-point quantization, 2) mixed-precision quantization and 3) attribution methods.

2.1 Fixed-point Quantization

Network quantization has aroused extensive interests in computer vision and machine learning due to the significant reduction in computation and storage complexity, and existing methods are divided into one-bit and multi-bit quantization. Binary networks constrain the network weights and activations in one bit at extremely high compression ratio. Hubara *et al.* (Hubara et al., 2016) and Courbariaux *et al.* (Courbariaux et al., 2016) replaced the multiply-add operations with xnor-bitcount via weight and activation binarization, and applied the straight-through estimators (STE) to differentiably optimize network parameters. Rastegari *et al.* (Rastegari et al., 2016) leveraged the scaling factor for weight and activation binarization to minimize the quantization errors. Liu *et al.* (Liu et al., 2018) added extra shortcut between consecutive convolutional layers to enhance the network capacity, they also presented the multinomial approximation of the sign function for accurate optimization. Wang *et al.* (Wang et al., 2019c) mined the channel-wise interactions to eliminate inconsistent signs in feature maps. Qin *et al.* (Qin et al., 2020) minimized the parameter entropy in inference and utilized the soft quantization in backward propagation to enhance the information retention. Since the performance gap between full-precision and binary networks is huge, multi-bit networks are presented for better accuracy-efficiency trade-off. Zhu (Zhu et al., 2016) trained an adaptive quantizer for network ternarization according to weight distribution. Gong *et al.* (Gong et al., 2019) applied the differentiable approximations for quantized networks to ensure the consistency between the gradient and the objective. Li *et al.* (Li et al., 2019) proposed the four-bit networks for object detection with hardware-friendly implementations, and overcome the training instabilities by custom batch normalization and outlier removal. However, the fixed-precision quantization ignores the redundancy variance across different layers and leads to suboptimal trade-off between accuracy and complexity in quantized networks.

2.2 Mixed-precision Quantization

The mixed-precision networks assign different bitwidths to weights and activations in various layers, which considers the redundancy variance in different components to obtain the optimal accuracy-complexity trade-off given hardware configurations. Existing mixed-precision quantization methods are mainly based on non-differentiable or differentiable search. For the former, Wang *et al.* (Wang et al., 2019a) presented a reinforcement learning model to learn the optimal bitwidth for weights and activations of each layer, where the model accuracy and complexity were considered in reward function. Wang *et al.* (Wang et al., 2020b) jointly searched the pruning ratio, the bitwidth and the architecture of the lightweight model from a hypernet via the evolutionary algorithms. Since the non-differentiable methods require huge search cost to obtain the optimal bitwidths, the differentiable search approaches are also introduced in mixed-precision quantization. Cai *et al.* (Cai and Vasconcelos, 2020) designed a hypernet where each convolutional layer consisted of parallel blocks in different bitwidths, and yielded the output by summing all blocks in various weights. Optimizing the block weight by back propagation and selecting the bitwidth with the largest value during inference achieved the optimal accuracy-complexity trade-off. Moreover, Yu *et al.* (Yu et al., 2020) further presented a barrier penalty to ensure that the searched models were within the complexity constraint. Yang *et al.* (Yang et al., 2020) decoupled the constrained optimization via Alternating Direction Method of Multipliers (ADMM), and Wang *et al.* (Wang et al., 2020c) utilized the variational information bottleneck to search the proper bitwidth and pruning ratio. Habi *et al.* (Habi et al., 2020) and Van *et al.* (van Baalen et al., 2020) directly optimized the quantization intervals for bitwidth selection of mixed-precision networks. However, differentiable search for mixed-precision quantization still needs a large amount of time due to the optimization of the large hypernet. In order to solve this, Dong *et al.* (Dong et al., 2019c), (Dong et al., 2019b) designed bitwidth assignment rules according to Hessian information. Nevertheless, the hand-crafted rules require expert knowledge and cannot adapt to the input data.

2.3 Attribution Methods

Attribution aims to produce human-understandable explanations for the predictions of neural networks. The contribution of each input component is calculated by examining its influence on the network output, which is displayed as the attribution in 2D feature maps. Early works (Erhan et al., 2009), (Simonyan et al., 2013), (Zhou et al., 2016) analyzed the sensitivity and the significance of each pixel by leveraging its gradients with respect to the objective optimization. Recent studies

on attribution extraction can be categorized into two types: gradient-based and relevance-based methods. For the first regard, Guided Backprop (Springenberg et al., 2014), Grad-Cam (Selvaraju et al., 2017) and integrated gradient (Sundararajan et al., 2017) combined the pixel gradients across different locations and channels for information fusion, so that more accurate attribution was obtained. Smoothgrad (Springenberg et al., 2014) proposed to smooth the derivatives of the class activation to eliminate the local fluctuation in sensitivity maps. For the latter regard, Zhang *et al.* (Zhang et al., 2018) constructed a hierarchical probabilistic model to mine the correlation between the input components and the prediction. Moreover, attribution was applied to enhance the generalizability in other computer vision tasks such as adversarial attack generation (Dong et al., 2019a). In this paper, we observe that the attribution rank consistency of feature maps between vanilla and compressed networks benefits downstream tasks for various data distribution, which is extended to generalizable mixed-precision quantization for significant search cost reduction.

3 Generalizable Mixed-Precision Quantization

In this section, we first introduce the mixed-precision quantization framework which suffers from significant search burden. Then we demonstrate the observation that the attribution rank consistency between quantized and full-precision models benefits visual analysis for various data distribution. Finally, we present the generalizable mixed-precision quantization via attribution rank preservation.

3.1 Preliminaries for Mixed-Precision Quantization

The goal of mixed-precision quantization is to search the proper bitwidth of each layer in order to achieve the optimal accuracy-complexity trade-off given hardware configurations. Since the distribution of the training and validation data for policy search significantly affects the acquired quantization strategy, existing methods require the training and validation data for quantization policy search and those for model deployment to come from the same dataset. However, the compressed models are usually utilized on large-scale datasets such as ImageNet, which causes heavy computational burden during quantization policy search. To address this, an ideal solution is to search for the quantization policy whose optimality is independent of the data distribution. Let $\boldsymbol{W}$ be the quantized network weight and $\mathcal{Q}$ be the quantization policy that assigns different bitwidths

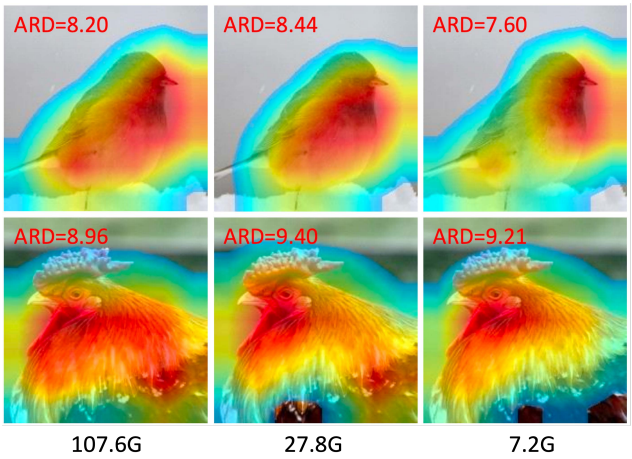


Fig. 3 The attribution of the mixed-precision networks in different BOPs with the optimal quantization policy. For the networks in low BOPs, the attribution is more concentrated although the rank remains similar. The concentrated attribution enables the model capacity to be sufficient by redundant attention removal, so that the promising performance is achieved.

to weights and activations in various layers. $\Omega(\mathcal{Q})$ means the computational complexity of the compressed networks with the quantization policy $\mathcal{Q}$. The search objective should be formulated in the following form:

$$\begin{aligned} \min_{\mathcal{Q}} \quad & \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{val}}\ \mathcal{L}(\boldsymbol{W}^*(\mathcal{Q}),\mathcal{Q},\boldsymbol{x}) \\ s.t. \quad & \boldsymbol{W}^*(\mathcal{Q}) = \arg\min \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{train}}\ \mathcal{L}(\boldsymbol{W},\mathcal{Q},\boldsymbol{x}) \\ & \Omega(\mathcal{Q}) \leqslant \Omega_0 \end{aligned} \tag{1}$$

where $\mathcal{L}(\boldsymbol{W},\mathcal{Q},\boldsymbol{x})$ represents the task loss for network weight $\boldsymbol{W}$, quantization policy $\mathcal{Q}$ and input $\boldsymbol{x}$. Ω_0 stands for the resource constraint of the deployment platform. In order to obtain the optimal mixed-precision networks, the quantization policy $\mathcal{Q}$ and the network weights $\boldsymbol{W}(\mathcal{Q})$ are alternatively optimized until convergence or the maximal iteration number. $\mathcal{D}_{val}$ depicts the dataset containing all validation images in deployment and $\mathcal{D}_{train}$ illustrates the dataset including given training images in bitwidth search, where the distribution gap between $\mathcal{D}_{val}$ and $\mathcal{D}_{train}$ may be sizable. Because $\mathcal{D}_{val}$ is intractable in realistic applications, it is desirable to find an alternative way to solve for the generalizable mixed-precision quantization policy.

3.2 Attribution Rank Consistency

Since acquiring all validation images in deployment is impossible, we solve the generalizable mixed-precision quantization policy via an alternative way. We observe that correctly locating the network attribution benefits visual analysis for various input data distribution. The

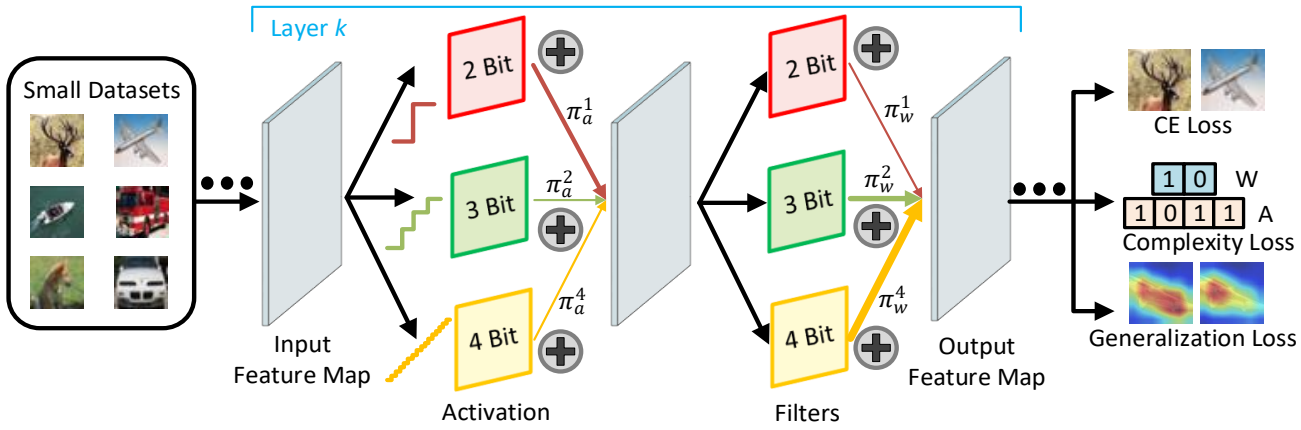


Fig. 4 The pipeline of our GMPQ. The hypernet consists of multiple parallel branches including filters and activations in different bitwidths. The output from various branches is added with learnable importance weights to construct output feature maps. Despite of the cross-entropy and complexity loss, we present additional generalization loss to optimize the network weights and branch importance weights, which enables the quantization policy searched on small datasets to be generalized on large-scale datasets. When the hypernet converges or achieves the maximum training epoch, we select the bitwidth represented by the branch with the largest importance weight to be the final quantization policy for finetuning.

feature attribution is formulated according to the loss gradient with respect to feature maps in the last layer, where the importance of the c_{th} feature map in the last convolutional layer for recognizing the objects from the t_{th} class is written as follows:

$$\alpha_c[t] = \frac{1}{Z}\sum_{m,n}\frac{\partial f(\boldsymbol{x})[t]}{\partial A_c[m,n]} \tag{2}$$

where $f(\boldsymbol{x})[t]$ means the output score for input $\boldsymbol{x}$ of the t_{th} class, and $A_c[m,n]$ represents the activation element in the m_{th} row and n_{th} column of the c_{th} feature map in the last convolutional layer. Z is a scaling factor that normalizes the importance into the range $[0,1]$. With the feature map visualization techniques presented in Grad-cam (Selvaraju et al., 2017), we obtain the feature attribution in the networks. We sum the feature maps from different channels with the attention weight calculated in (2), and remove the influence from opposite pixels via the ReLU operation. The feature attribution in the last convolutional layer with respect to the t_{th} class is formulated in the following:

$$M[t] = ReLU(\sum_{c}\alpha_c[t]\cdot\boldsymbol{A_c}) \tag{3}$$

The feature attribution only preserves the supportive features for the given class, and the negative features related to other classes are removed. Meanwhile, the attribution is upsampled with the size of input images by bilinear interpolation (Selvaraju et al., 2017) in order to keep the resolution consistency. Preserving attribution similarity have been proven to be meta knowledge that can be transferred among different data distribution in many tasks including adversarial attack Dong et al. (2019a); Wang et al. (2021); Wu et al. (2020) and domain adaptation Gao et al. (2023); Wang et al. (2019b); Zunino et al. (2021), and extensive experiments have also verified that they can enhance the model performance on novel data. Please refer to Appendix J for more detailed formulation. Therefore, we expect similar attribution between quantized and full-precision models to enhance the generalization ability of the searched quantization policies.

The full-precision networks achieve high performance due to paying more attention to important parts in the image, while the quantized models deviate the attribution from that of the full-precision networks due to the limited capacity. Figure 3 demonstrates the attribution of networks with the optimal quantization policy in different complexity, where attribution of networks in lower capacity is more concentrated due to the limited carried information. As the network capacity gap between the quantized networks and their full-precision counterparts is huge, directly enforcing the attribution consistency fails to remove the redundant attention in the compressed model, which causes capacity insufficiency with performance degradation. Therefore, we preserve the attribution rank consistency between the quantized networks and their full-precision counterparts for generalizable mixed-precision quantization policy search. The attribution rank illustrates the importance order of different pixels for model predictions. Constraining attribution rank consistency enables the quantized networks to focus on important regions, which adaptively adjusts the attribution distribution without capacity insufficiency.

3.3 Generalizable Mixed-Precision Quantization via Attribution Rank Preservation

Our GMPQ can be leveraged as a plug-and-play module for both non-differentiable and differentiable search methods. Since differentiable methods achieve the competitive accuracy-complexity trade-off compared with non-differentiable approaches, we employ the differentiable search framework (Cai and Vasconcelos, 2020; Yang et al., 2020; Yu et al., 2020) to select the optimal mixed-precision quantization policy. We design a hypernet with N_a^k and N_w^k parallel branches for convolution filters and feature maps in the k_{th} layer. N_a^k and N_w^k represent the size of the search space for weight and activation bitwidths. The parallel branches are assigned with various bitwidths whose output is summed with the configuration parameters for weight and activation respectively to form the intermediate feature maps. Figure 4 depicts the pipeline of our GMPQ. The feed-forward propagation for each layer in the K-layer hypernet is written as follows:

$$\boldsymbol{z}^k = \sum_{i=1}^{N_w^k} \pi_{w,i}^k \boldsymbol{w}_i^k (\sum_{j=1}^{N_a^k} \pi_{a,j}^k \boldsymbol{a}_j^k)$$
$$s.t. \sum_{i=1}^{N_w^k} \pi_{w,i}^k = 1, \quad \sum_{i=1}^{N_a^k} \pi_{a,i}^k = 1, \quad \pi_{w,i}^k, \pi_{a,i}^k \in [0,1], \quad (4)$$

where $\boldsymbol{z}^k$ means the output intermediate feature maps of the k_{th} layer. $\boldsymbol{a}_j^k$ represents the output of the j_{th} activation quantization branch in the k_{th} layer, and $\boldsymbol{w}_i^k$ is the i_{th} convolution filter of the k_{th} layer. $\pi_{a,i}^k$ and $\pi_{w,i}^k$ stand for the configuration parameters for the i_{th} quantized activation and filter branch in the k_{th} layer, which are normalized in the following way:

$$\pi_{w,i}^k = \tilde{\pi}_{w,i}^k / \sum_{i=1}^{N_w^k} \tilde{\pi}_{w,i}^k, \qquad \pi_{a,i}^k = \tilde{\pi}_{a,i}^k / \sum_{i=1}^{N_a^k} \tilde{\pi}_{a,i}^k \qquad (5)$$

where $\tilde{\pi}_{w,i}^k$ and $\tilde{\pi}_{a,i}^k$ mean the learnable coefficients before normalization, and are optimized by the gradients of the hypernet.

As we observe that the attribution rank consistency between quantized networks and their full-precision counterparts enables the compressed models to possess the discriminative power of the vanilla model regardless of the data distribution, we impose the attribution rank consistency constraint in optimal quantization policy search despite of the accuracy and efficiency objective. In order to obtain the optimal accuracy-complexity trade-off for generalizable mixed-precision quantization, the learning objective is formulated in the Lagrangian form:

$$\mathcal{R} = \mathcal{R}_E(\boldsymbol{W}, \mathcal{Q}, \boldsymbol{x}) + \zeta \mathcal{R}_C(\mathcal{Q}) + \eta \mathcal{R}_G(\boldsymbol{W}, \mathcal{Q}, \boldsymbol{x}) \quad (6)$$

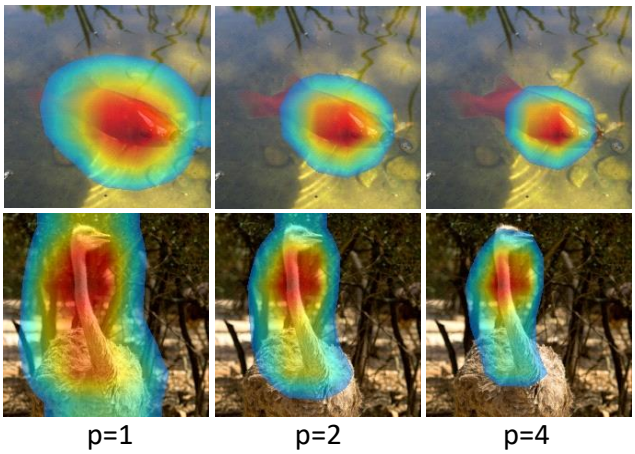


Fig. 5 The L_p norm of the attribution for the full-precision networks with different p. The attribution is more concentrated for larger p while the rank keeps same.

where $\mathcal{R}_E(\boldsymbol{W}, \mathcal{Q}, \boldsymbol{x})$, $\mathcal{R}_C(\mathcal{Q})$ and $\mathcal{R}_G(\boldsymbol{W}, \mathcal{Q}, \boldsymbol{x})$ respectively mean the task, complexity and the generalization risk for the networks with weight $\boldsymbol{W}$ and quantization policy $\mathcal{Q}$ for the input $\boldsymbol{x}$. ζ and η are the hyperparameters to balance the importance of the complexity risk and generalization risk in the overall learning objective. In differentiable policy search, $\mathcal{R}_E(\boldsymbol{W}, \mathcal{Q}, \boldsymbol{x})$ is represented by the objective of vision tasks, and $\mathcal{R}_C(\mathcal{Q})$ is defined as the expected Bit-operations (BOPs) (Bethge et al., 2020; Cai and Vasconcelos, 2020; Wang et al., 2020c):

$$\mathcal{R}_C(\mathcal{Q}) = \sum_{k=1}^{K} (\sum_{i=1}^{N_w^k} \pi_{w,i}^k q_{w,i}^k) \cdot (\sum_{i=1}^{N_a^k} \pi_{a,i}^k q_{a,i}^k) \cdot B_{full}^k \quad (7)$$

where $q_{w,i}^k$ and $q_{a,i}^k$ stand for the bitwidth of the i_{th} branch of weights and activations in the k_{th} layer, and B_{full}^k means the BOPs of the k_{th} layer in the full-precision network. K represents the number of layers of the quantized model. As the attribution rank consistency between the full-precision networks and their quantized counterparts enhances the generalizability of the mixed-precision quantization policy, we define the generalization risk in the following form:

$$\mathcal{R}_G = \sum_{i,j} ||r(M_{q,ij}[y_x]) - r(M_{f,ij}[y_x])||_2^2 \qquad (8)$$

where $M_{q,ij}[y_x]$ represents the pixel attribution in the i_{th} row and j_{th} column of the input images with respect to the class y_x in the quantized networks, and $M_{f,ij}[y_x]$ demonstrates the corresponding variable in full-precision models. y_x means the label of the input $\boldsymbol{x}$, and $||\cdot||_2$ is the element-wise l_2 norm. $r(\cdot)$ stands for the attribution rank, which equals to k if the element is the k_{th} largest in the attribution map. We only preserve

the attribution rank consistency for top-k pixels with the highest attribution in the full-precision networks, as low attribution is usually caused by noise without clear information. Since minimizing the generalization risk is NP-hard, we present the capacity-aware attribution imitation to differentiably optimize the objective.

To benefit the objective optimization process, we normalized the original attribution score acquired from Grad-cam by dividing summation over all pixel values. We enforce attribution of the mixed-precision networks to approach the l_p norm of that in full-precision models, because the l_p norm preserves the rank consistency while adaptively selects the attribution distribution according to the network capacity. The generalization risk is rewritten as follows for efficient optimization:

$$\mathcal{R}_G = \sum_{i,j} ||M_{q,ij}[y_x] - \frac{M_{f,ij}[y_x]^p}{\sum_{i,j} M_{f,ij}[y_x]^p}||_2^2 \quad (9)$$

Large p leads to concentrated attribution and vice versa, as shown in Figure 5. Excessively large p enforces quantized models to focus on small regions that fail to fully utilize the capacity, and extremely small p leads to paying attention to large area with capacity insufficiency. Therefore, we adjust the attribution concentration according to the network capacity with hyperparameters Q_w^0 and Q_a^0 for L-layer networks, where quantized networks in low bitwidths can select more concentrated attribution and vice versa:

$$p = \frac{1}{L}\sum_{k=1}^{L}(Q_w^0 / \sum_{i=1}^{N_w^k} \pi_{w,i}^k q_{w,i}^k) \cdot (Q_a^0 / \sum_{i=1}^{N_a^k} \pi_{a,i}^k q_{a,i}^k) \quad (10)$$

Since the task, complexity and generalization risks are all differentiable, we optimize the hypernet weights and the branch importance weights iteratively in an end-to-end manner. When the hypernet converges or achieves the maximum training epoch, the bitwidth represented by the branch with the largest important weight is selected to form the final quantization policy. We finetune the quantized networks with the data in deployment to acquire the final model applied in realistic applications. GMPQ searches quantization policies on small datasets with generalization constraint, which leads to high performance on large-scale datasets in deployment with reduced search cost.

4 Robust Generalizable Mixed-Precision Quantization

We first propose hierarchical partitioning based attribution smoothing, and then present the dynamic capacity-aware attribution imitation to optimally adjust the attribution distribution according to the sample hardness for each input.

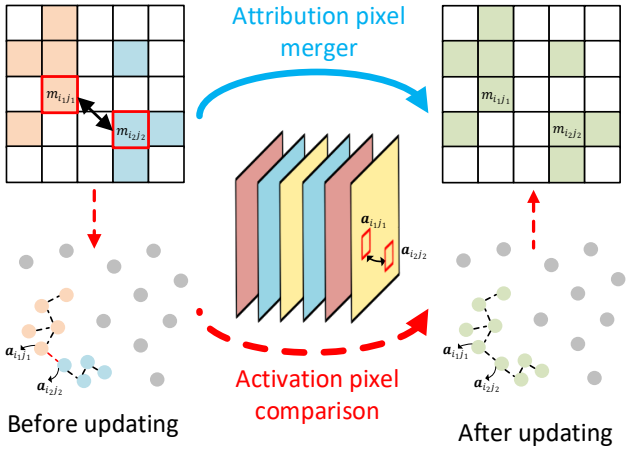


Fig. 6 The attribution partition process, where different colors in attribution maps represent various partitions. The attribution pixels $m_{i_1j_1}$ and $m_{i_2j_2}$ are defined to be semantically similar if the distance between their counterparts $\boldsymbol{a}_{i_1j_1}$ and $\boldsymbol{a}_{i_2j_2}$ in activation maps is no more than the threshold, and the partitions they belong to are merged with the same index assignment (the top row). Therefore, pixels in both partitions are merged into one partition (the bottom row).

4.1 Attribution Smoothing via Hierarchical Partitioning

The attribution ranks in full-precision models usually contain significant errors because slight attribution noise is amplified by ranking operations, which fails to reveal the true region importance. While GMPQ enforces the quantized networks to mimic attribution ranks of the full-precision counterparts for each pixel, the attribution rank errors in full-precision networks hinder the quantized models to correctly locate the attribution. In order to address these limitations, we propose robust generalizable mixed-precision quantization (R-GMPQ) to eliminate attribution rank errors, where attribution in full-precision models is smoothed via hierarchical partitioning.

Attribution pixels sharing similar semantics should be assigned with the same attribution rank, where the semantic similarity can be evaluated by the pixel distance of activation maps in the last layer (Park et al., 2020; Zagoruyko and Komodakis, 2016). As attribution is upsampled with the size of input images, the activation maps are also re-scaled to the same resolution for semantic similarity measurement by bilinear interpolation. In order to correctly locate the region importance without significant rank errors, we assign attribution pixels sharing similar semantics with the same rank. Inspired by techniques presented in feature denoising (Du et al., 2016; Huang et al., 2016; Xie et al., 2019), we smooth the attribution via partition statistics so that rank errors caused by noise are alleviated.

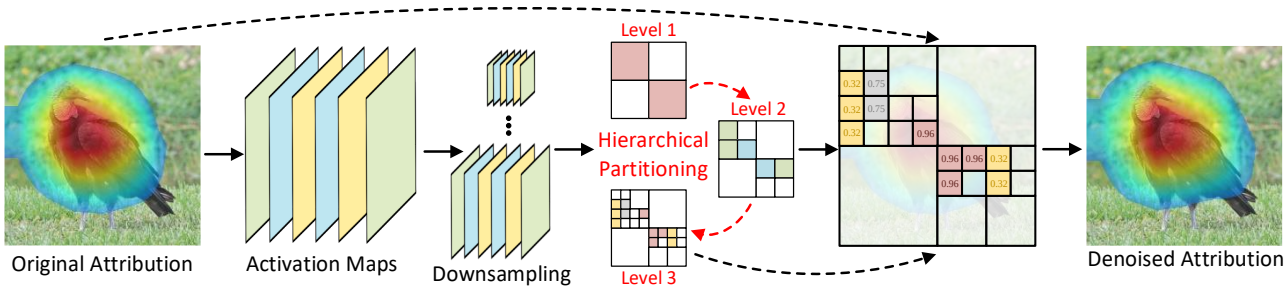


Fig. 7 The pipeline of hierarchical attribution partitioning. The re-scaled activation maps is first downsampled into L gradient maps with different resolution. Then we explore the semantically similar pixels from top activation maps to the bottom ones, where we only search the pixels in the region whose downsampled counterparts in the previous level are semantically similar. Finally, we smooth each attribution pixel with the statistics in the partition. In this figure, the semantic similarity evaluation within each blank area during hierarchical partitioning is omitted for representation simplicity.

Let us denote the attribution pixel in the i_{th} row and j_{th} column as m_{ij}, whose counterpart in the activation map is $\boldsymbol{a}_{ij} \in \mathbb{R}^{1\times c}$ with c channels. Two attribution pixels $m_{i_1j_1}$ and $m_{i_2j_2}$ share similar semantics if the distance of their counterparts on the activation maps is no more than the threshold, and the partitions that the attribution pixels belong to should be merged with the same attribution value assignment to keep the rank consistency. Therefore, the activation pixels share the same division arrangement with the attribution pixels. By enumerating activation pixel pairs, we construct the attribution partitions for different images, where the update for attribution partition in each step of the enumeration is implemented as follows:

$$\begin{aligned}\mathcal{T}^*_{i_1j_1} &= \mathcal{T}_{i_1j_1} \cup \mathcal{T}^I_{i_2j_2}(||\boldsymbol{a}_{i_1j_1} - \boldsymbol{a}_{i_2j_2}|| \leqslant \gamma\sigma_a)\\ \mathcal{T}^*_{i_2j_2} &= \mathcal{T}_{i_2j_2} \cup \mathcal{T}^I_{i_1j_1}(||\boldsymbol{a}_{i_1j_1} - \boldsymbol{a}_{i_2j_2}|| \leqslant \gamma\sigma_a)\\ i_1, i_2 \in \{1, 2, ..., h\}, &\qquad j_1, j_2 \in \{1, 2, ..., w\}\end{aligned} \tag{11}$$

where $\mathcal{T}^*_{ij}$ and $\mathcal{T}_{ij}$ represent the partition that $\boldsymbol{a}_{ij}$ belongs to after and before update in each step of the enumeration. $\mathcal{T}^I_{ij}(x)$ stands for the the partition $\mathcal{T}_{ij}$ for true x and means the empty set otherwise. γ is a hyperparameter and σ_a means the standard deviation of the activation pixel norm. Meanwhile, h and w demonstrate the height and width of the activation maps. Figure 6 depicts the attribution partition process. Since the standard deviation demonstrates the fluctuation of activation pixels, assigning the same partition index to attribution pixels where the distance between their counterparts in activation maps is less than $\gamma\sigma_a$ enables adaptive exploration of pixel semantic similarity.

However, directly enumerating the pixel pairs in activation maps leads to $O(h^2w^2c^2)$ computational complexity, which significantly increases the cost for mixed-precision quantization policy search. In order to efficiently evaluate the semantic similarity among all attribution pixels, we present the hierarchical partitioning strategy that decomposes the overall semantic similarity computation into different levels, where pairwise semantic similarity is sparsely evaluated in each hierarchy. Therefore, the computational complexity of semantic similarity measurement is low in each level, and combining the obtained similarity across various hierarchies yields the overall semantic consistency efficiently. We downsample the activation maps in multiple hierarchies with different resolution by average pooling, where activation maps in the l_{th} level with the spatial resolution $h_l \times w_l$ is denoted as $\boldsymbol{a}^l \in \mathbb{R}^{h_l\times w_l\times c}$. We define that activation maps in the first level are in the lowest resolution and vice versa, where semantics are represented in various hierarchies. Therefore, semantically similar pixels in previous level indicate that the corresponding upsampled region of activations in the current hierarchy contain semantically similar pixels. More specifically, we only evaluate the similarity among activation pixels in the region whose downsampled counterparts in the previous level are in the same partition. The partitions of activation maps in the l_{th} hierarchy are constructed via the following way:

$$\begin{aligned}\mathcal{T}^{l*}_{i_1j_1} &= \mathcal{T}^l_{i_1j_1} \cup \mathcal{T}^{lI}_{i_2j_2}(||\boldsymbol{a}^l_{i_1j_1} - \boldsymbol{a}^l_{i_2j_2}|| \leqslant \gamma\sigma^l_a)\\ \mathcal{T}^{l*}_{i_2j_2} &= \mathcal{T}^l_{i_2j_2} \cup \mathcal{T}^{lI}_{i_1j_1}(||\boldsymbol{a}^l_{i_1j_1} - \boldsymbol{a}^l_{i_2j_2}|| \leqslant \gamma\sigma^l_a)\\ (i_i, j_1) \in \{(1,1), ..., (h_l, w_l)\}, &\quad (i_2, j_2) \in \mathcal{I}^l_{i_1j_1}\end{aligned} \tag{12}$$

where $\boldsymbol{a}^l_{ij}$ represents the pixel in the i_{th} row and j_{th} column of activation maps in the l_{th} level, and σ^l_a means standard deviation of the pixel norm in activations of the l_{th} level. $\mathcal{T}^{m*}_{ij}$ and $\mathcal{T}^m_{ij}$ represent the partition of $\boldsymbol{a}^m_{ij}$ after and before update in each step during the enumeration, $\mathcal{T}^{mI}_{ij}(x)$ stands for the the partition $\mathcal{T}^m_{ij}$ for true x and means the empty set otherwise. $\mathcal{I}^l_{ij}$ demonstrates the activation regions in the l_{th} level whose downsampling pixels in the previous hierarchy are semantically similar with those of $\boldsymbol{a}^l_{ij}$. Figure 7 demonstrates the pipeline of our hierarchical partitioning for attribution smoothing. By sparsely exploring the semantic similarity among activation pixels across different hierarchies, we efficiently partition the attribution according to the activation divisions in the final level.

Algorithm 1 R-GMPQ

Input: Full-precision network $\mathcal{N}_f$, hierarchy level L, resolution reduction ratio R.
Output: Mixed-precision network $\mathcal{N}_q$.
for $l = 1, 2, ..., L$ **do**
Pooling original activation with kernel size R^l to acquire downsampled counterparts $\boldsymbol{a}^l$.
Enumerating each pixel in attribution for partition assignment via (12).
end for
Assigning the smoothed value of attribution via (13).
Optimizing the supernetwork with (6).
Discretizing the supernetwork by selecting the bitwidth with the largest important weights.
return The finetuned quantized network $\mathcal{N}_q$.

For attribution smoothing in each division, we assign all pixels with the mean value of attribution in the partition, and the attribution pixel m_{ij} of full-precision networks is denoised as follows:

$$m^*_{ij} = \frac{1}{|S(m_{ij})|} \sum_{m_{xy} \in S(m_{ij})} m_{xy} \tag{13}$$

where m^*_{ij} is the denoised counterpart of m_{ij}. $S(m_{ij})$ means attribution pixels in the same division with m_{ij}, whose number of pixels are represented by $|S(m_{ij})|$. Smoothing attribution maps according to semantic similarity among pixels alleviates rank errors, so that the correct region importance is revealed in the attribution of full-precision networks for quantized models to imitate. Since the attribution smoothing is only required once before the generalizable mixed-precision quantization policy search, the computational cost to obtain the optimal compression strategy is negligible.

4.2 Dynamic Capacity-aware Attribution Imitation

The optimal attribution concentration degree varies for input samples in different hardness given the quantization policy, where the same concentration degree leads to insufficient model capacity for hard images and fails to fully utilize the network capacity for easy ones during attribution imitation. Since the fixed capacity-aware attribution imitation presented in GMPQ ignores the sample hardness variation, the distribution of acquired attribution in quantized networks is usually over-concentrated for easy samples and excessively divergent for hard ones. Figure 8 demonstrates the attribution in the optimally quantized networks and the l_p norm of the attribution in full-precision models. The optimally quantized model locates the attribution more similarly to the l_p norm of attribution in full-precision networks with larger p for harder samples.

The capacity of full-precision networks is regarded to be sufficient, whose attribution rank is mimicked by

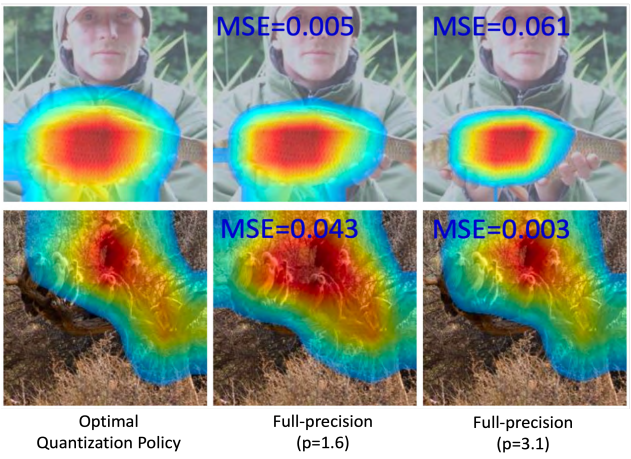


Fig. 8 The attribution in optimally quantized models and the l_p norm of attribution in full-precision counterparts for samples in various hardness. The mean squared errors (MSE) of the l_p norm between full-precision and quantized attribution are also demonstrated. The quantized networks locate the attribution of easy images (the top row) similarly to the l_p norm of attribution in full-precision models with smaller p, where images containing more objects and more complex background (the bottom row) usually requires more concentrated attribution due to the higher hardness.

quantized models. Therefore, we utilize the division of the task risk between quantized and full-precision networks to evaluate the capacity insufficiency caused by quantization. In order to dynamically choose the optimal attribution distribution for input in various hardness, our dynamic capacity-aware attribution imitation (DCAI) employs the following the generalization risk:

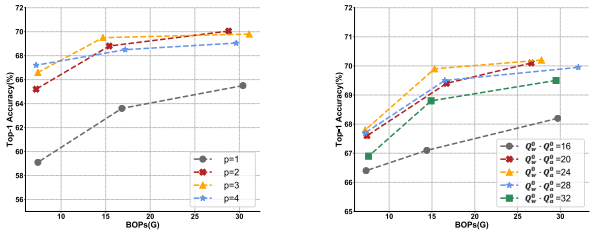
$$\mathcal{R}_G = \sum_{i,j} ||M_{q,ij}[y_x] - \frac{M^*_{f,ij}[y_x]^{\mathcal{R}^q_E/\mathcal{R}^f_E}}{\sum_{i,j} M^*_{f,ij}[y_x]^{\mathcal{R}^q_E/\mathcal{R}^f_E}}||^2_2 \tag{14}$$

where $\mathcal{R}^q_E$ and $\mathcal{R}^f_E$ represent task loss for quantized and full-precision models respectively, and $M^{*k}_{f,ij}[y_x]$ stands for the denoised attribution of full-precision networks. The presented DCAI strengthens the generalizability of the mixed-precision quantization policy due to the dynamic attribution concentration degree that fully utilizes the model capacity without insufficiency for samples in different hardness. Algorithm 1 shows the pseudo code of the overall R-GMPQ pipeline.

5 Experiments

In this section, we conducted extensive experiments to evaluate our methods on ImageNet for image classification and on Pascal VOC and COCO for object detection. We first introduce the implementation details of our GMPQ and R-GMPQ. In the following ablation study, we then evaluated the influence of value

assignment strategy for p in the capacity-aware attribution imitation, investigated the effects of different terms in the overall risk function and discovered the impact of datasets for quantization policy search. Moreover, we also empirically analyze the effectiveness and efficiency of the hierarchical attribution partitioning, and show the superiority of the dynamic capacity-aware attribution imitation on performance. Finally, we compare our GMPQ and R-GMPQ with the state-of-the-art mixed-precision networks with respect to accuracy, model complexity, the compression ratio and search cost. The search cost represents the computation resource consumption measured by GPU hours to acquire the mixed-precision quantization policy, and the compression ratio is defined as the ratio between the BOPs of quantized networks to those of full-precision counterparts.



(a) Fixed strategy (b) Capacity-aware strategy

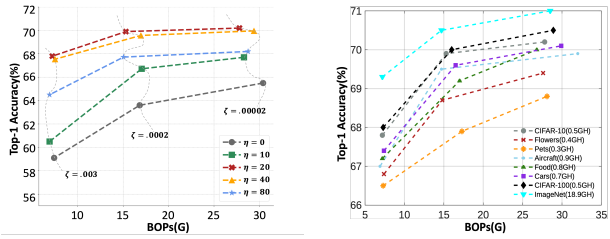
Fig. 9 The accuracy-complexity trade-off of (a) fixed and (b) capacity-aware value assignment strategies for p in (10), where hyperparameters were also varied.

5.1 Implementation Details

For mixed-precision network deployment, we evaluated the quantized networks on ImageNet for image classification and on PASCAL VOC and COCO for object detection. ImageNet (Deng et al., 2009) approximately contains 1.2 billion and 50k images for training and validation from $1,000$ categories. For training, 224×224 random region crops were applied from the resized image whose shorter side was 256. During the inference stage, we utilized the 224×224 center crop. The PASCAL VOC dataset (Everingham et al., 2010) collected images from 20 categories, where we fintuned our mixed-precision networks on VOC 2007 and VOC 2012 trainval sets containing about 16k images and tested our GMPQ and R-GMPQ on VOC 2007 test set consisting of 5k samples. Following (Everingham et al., 2010), we used the mean average precision (mAP) as the evaluation metric. The COCO dataset consists of images from 80 different categories, and our experiments were conducted on the 2014 COCO object detection track. We trained our model with the combination of 80k images from the training set and 35k images selected from validation set (trainval35k (Bell et al., 2016)), and tested our method on the remaining minival validation set (Bell et al., 2016) including 5k images. Following the standard COCO evaluation metric (Lin et al., 2014), we apply the mean average precision (AP) for IoU $\in [0.5 : 0.05 : 0.95]$ as the evaluation metric. We also report average precision with the IOU threshold 50% and 75% represented as AP_{50} and AP_{75} respectively. Moreover, the average precision of small, medium and large objects notated as AP_s, AP_m and AP_l are also depicted.

We trained our GMPQ with MobileNet-V2 (Sandler et al., 2018), ResNet18 and ResNet50 (He et al., 2016), DeiT-S/T Touvron et al. (2021) architectures for image classification, and applied VGG16 (Simonyan and Zisserman, 2014) with SSD framework (Liu et al., 2016) and ResNet18 with Faster R-CNN (Ren et al., 2015) for object detection. The bit-width in the search space for network weights and activations is 2-8 bit for MobileNet-V2 and 2-4 bit for other architectures. Inspired by (Cai and Vasconcelos, 2020), we utilized compositional convolution whose filters were weighted sum of each quantized filters in different bitwidths, so that complex parallel convolution was avoided. We updated the importance weight of different branches and the network parameters simultaneously. The hyperparameters Q_w^0 and Q_a^0 in capacity-aware attribution imitation were set to 4 and 6 respectively. Meanwhile, we only minimize the distance between attribution in quantized networks and l_p norm of that in full-precision counterparts for top-500 pixels with the highest attribution in the real-valued model. We downsampled the activation maps with three levels in the hierarchical partitioning, where the height and width both decrease by two times between adjacent levels. Meanwhile, the hyperparameter γ that controls the attribution partition merger in (12) was assigned to 0.05.

For evaluation on ImageNet, we finetuned the mixed-precision networks with the Adam (Kingma and Ba, 2014) optimizer. The learning rate started from 0.001 and decayed twice by multiplying 0.1 at the 20_{th} and 30_{th} epoch out of the total 40 epochs. For object detection, the backbone was pretrained on ImageNet and then finetuned on PASCAL VOC and COCO with the same hyperparameter settings. The batchsize was set to 256 in all experiments. By adjusting the hyperparameters ζ and η in (6), we obtained the mixed-precision networks at different accuracy-complexity trade-offs.



(a) Varying ζ and η (b) Varying datasets

Fig. 10 (a) The accuracy-complexity trade-off for different η, where ζ was varied to select various network capacity. (b) The top-1 accuracy on ImageNet, the BOPs and the average search cost of the mixed-precision quantization policy searched on different small datasets, where GH means GPU hours for the search cost.

5.2 Ablation Study

In this section, we analyze the effect of attribution rank preservation and the capacity-aware attribution imitation in GMPQ at first, and then show the effectiveness of the proposed hierarchical attribution partitioning and dynamic capacity-aware attribution imitation in R-GMPQ by the ablation study. Unless clarified, we compressed the ResNet18 architecture whose mixed-precision quantization policy was searched on CIFAR-10 and evaluated on ImageNet for all experiments in this section.

5.2.1 Ablation Study for GMPQ

In order to investigate the effectiveness of attribution rank preservation, we assign the value of p in CAI with different strategies. By varying the hyperparameters ζ and η in the overall risk (6), we evaluated the influence of task, complexity and generalization risks with respect to the model accuracy and efficiency. Moreover, we searched the generalizable mixed-precision quantization policy on different small datasets to discover the effects on accuracy-complexity trade-offs and search cost.

Effectiveness of value assignment strategies for p: To investigate the influence of value assignment strategies for p on the accuracy-complexity trade-off, we searched the mixed-precision quantization policy with fixed and capacity-aware p value. For fixed p, we set the value as 1, 2, 3 and 4 that constrains the attribution of quantized networks with various concentration degree. The capacity-aware strategy assigns p with the strategy shown in (10), where the product of Q_w^0 and Q_a^0 was varied in the ablation study. Figure 9(a) and 9(b) demonstrate the accuracy-complexity trade-off for fixed and capacity-aware value assignment strategies of p respectively with different hyperparameters. The optimal accuracy-complexity curve in capacity-aware strategy outperforms that in fixed strategy, which indicates the importance of attribution variation with respect to network capacity. For fixed strategy, medium p outperforms other values. Small p causes model capacity insufficiency for quantized networks and large p fails to utilize the network capacity. For capacity-aware strategy, setting the product of Q_w^0 and Q_a^0 to 24 results in the optimal accuracy-complexity trade-off.

Influence of hyperparameters in overall risk (6): In order to verify the effectiveness of the generalization risk, we report the performance with different η. Meanwhile, we also varied the hyperparameter ζ to obtain different accuracy-complexity trade-offs. Figure 10(a) illustrates the results, where medium η achieves the best trade-off curve. Large η fails to leverage the supervision from annotated labels, and small η ignores the attribution rank consistency which enhances the generalization ability of the mixed-precision quantization policy. With the increase of ζ, the resulted policy prefers lightweight architectures and vice versa. For different η, the same assignment of ζ selects similar BOPs in the accuracy-complexity trade-off.

Effects of datasets for network quantization policy search: We searched mixed-precision quantization policy on different small datasets including CIFAR-10, Cars, Flowers, Aircraft, Pets, Food, CIFAR-100 to discover the effects on model accuracy and efficiency. We also provide the performance of quantization policies searched on the original ImageNet for comparison. Figure 10(b) demonstrates the top-1 accuracy and the BOPs for the optimal mixed-precision networks obtained on different small datasets. We also show the average search cost across all computation cost constraint in the legend, where GH means GPU hours that measures the search cost. Among all small datasets, the mixed-precision networks searched on CIFAR-100 achieves the best accuracy-efficiency trade-off, because the size of CIFAR-100 is the largest with the diverse categories. Moreover, the gap of object category between CIFAR-100 and ImageNet is the smallest compared with other datasets. Leveraging extremely small numbers of data for policy search may lead to overfitting, and utilizing the full training set of small datasets is required to search generalizable mixed-precision quantization strategies without the risk of overfitting.

5.2.2 Ablation Study for R-GMPQ

Since the hierarchical attribution partitioning efficiently smooths the attribution with similar semantics, the rank errors are alleviated and the true region importance is revealed by the attribution ranks of full-precision models. To investigate the influence of the hierarchy set-

Table 2 The BOPs (G), top-1 accuracy and the computational cost (hours) of hierarchical attribution partition w.r.t. different numbers of hierarchies, resolution reduction ratios between adjacent levels and the assignment of the hyperparameter γ. RR ratio means the resolution reduction ratio between adjacent levels.

Hierarchies	RR ratio	$\gamma = 0.01$			$\gamma = 0.05$			$\gamma = 0.1$			$\gamma = 0.5$		
		BOPs	Top-1	Cost	BOPs	Top-1	Cost	BOPs	Top-1	Cost	BOPs	Top-1	Cost
One-level	-	7.1	68.3	8.54	7.2	68.6	8.25	7.2	68.4	8.60	7.5	68.2	8.72
	2	7.1	61.2	0.22	7.2	61.0	0.22	7.3	60.7	0.21	7.3	60.5	0.19
	4	7.4	60.5	0.09	7.3	60.7	0.10	7.2	59.9	0.10	7.6	60.1	0.09
Two-level	2	7.3	68.2	0.68	7.4	68.5	0.62	7.2	68.4	0.66	7.5	68.2	0.69
	4	7.2	68.1	0.26	7.2	68.4	0.24	7.6	68.3	0.27	7.2	68.0	0.49
Three-level	2	7.1	68.2	0.20	7.2	68.5	0.15	7.5	68.3	0.16	7.3	68.1	0.41
	4	7.2	68.0	0.09	7.4	68.3	0.07	7.1	68.1	0.08	7.4	67.8	0.18
Five-level	2	7.5	67.5	0.07	7.3	68.1	0.04	7.2	67.8	0.05	7.4	67.1	0.13

Table 3 The model storage cost (M), model computational cost (G), top-1 accuracy (%) and search cost (GPU hours) on ImageNet. Param. means the model storage cost, and Comp. stands for the compression ratio of BOPs.

Methods	Param.	BOPs	Comp.	Top1	Cost
CAI	4.1	7.3	254.9	68.0	0.91
DCAI	3.5	7.2	258.5	68.5	0.95

Table 4 The average attribution rank difference across all test images for different mixed-precision quantization methods.

	ImageNet	VOC	COCO
Random	39.94	40.07	78.95
GMPQ	18.14	19.63	38.21
R-GMPQ	13.57	16.59	30.33

tings on the smoothing efficiency and the policy generalizability, we implemented the hierarchical attribution partitioning with various numbers of activation levels and different resolution reduction ratio for adjacent levels. Besides, the impact of the hyperparameter γ that controls the activation partition merger was also explored. Meanwhile, DCAI adaptively adjusts p in l_p norm of the full-precision model attribution based on the sample hardness, so that the capacity of quantized networks is fairly evaluated and the optimal attribution concentration degree is dynamically selected in the imitation for each input. To verify the effectiveness of DCAI, our methods with static capacity-aware and dynamic capacity-aware strategy of p are compared with respect to the accuracy-complexity trade-off and the search cost. Finally, we compare the ARD of GMPQ and R-GMPQ to show the effectiveness of attribution smoothing and DCAI in attribution rank consistency preservation.

Impacts of activation levels and resolution reduction ratio between adjacent levels: We partitioned the attribution maps with various numbers of levels and resolution reduction ratios between adjacent levels for smoothing, where the accuracy-complexity trade-off and the computational cost for partitioning is demonstrated in Table 2. Increasing the number of hierarchies reduces the computational cost for partitioning while degrades the policy generalizability, since decomposing the search space to more subgroups decreases the solution optimality. However, the attribution partitioning less than three levels is not sensitive to the number of hierarchies, and we choose the three-level hierarchies for attribution partitioning to achieve high efficiency. Meanwhile, higher reduction ratio also results in lower computational cost for attribution partitioning and worse accuracy-complexity trade-off on ImageNet, because aggressive activation downsampling leads to significant information loss. Therefore, we assigned the resolution reduction ratio with two in other experiments.

Influence of the hyperparameter γ in (12): The hyperparameter γ controls the activation division merger in hierarchical partitioning. Table 2 shows the accuracy-complexity trade-off and the computational time of attribution partitioning with various γ, where the medium γ achieves the optimal trade-off. Large γ smooths excess pixels by the statistics with significant information loss, and small γ fails to alleviate the discretization errors due to the biased statistics of insufficient pixels for each partition.

Comparison between CAI and DCAI: We compare static capacity-aware and dynamic capacity-aware strategies for value assignment of p in order to show the superiority of DCAI, where the accuracy-complexity is shown in Table 3. Therefore, we draw the conclusion that DCAI can further strengthen the policy generalizability with negligible extra search cost.

ARD of GMPQ and R-GMPQ: We report the ARD of quantized models with policies searched with our GMPQ and R-GMPQ on CIFAR-10 in Table 4, where different quantized networks evaluated on the same dataset have similar model complexity. The experimental results prove that attribution rank consistency learned from one dataset can be generalized to different data distribution. Compared with the random policies, our method can significantly reduce the rank inconsis-

Table 5 The top-1 accuracy (%) on ImageNet, parameter storage cost (M), model computational cost (G) and the search cost (GPU hours). Param. means the model storage cost, and Comp. stands for the compression ratio of BOPs. The training cost for ResNet18, ResNet50 and MobileNet-V2 is 60.8, 80.9 and 37.4 GPU hours.

Methods	Param.	BOPs	Comp.	Top1	Cost.
ResNet18					
Baseline	46.8	1853.4	–	69.7	–
ALQ	1.8	58.5	31.7	67.7	34.7
HAWQ	5.8	34.0	54.5	68.5	15.6
GMPQ	5.4	27.8	66.7	70.2	0.5
R-GMPQ	5.3	27.1	68.3	70.4	0.6
APoT	4.6	16.3	113.8	69.8	–
SDQ	5.2	15.7	118.1	69.1	9.0
GMPQ	4.1	15.3	121.0	69.9	0.6
R-GMPQ	3.8	15.6	118.7	70.1	0.7
EdMIPS	4.7	7.2	258.0	65.9	9.5
EdMIPS-C	4.5	7.4	251.9	59.1	0.6
GMPQ	3.7	7.2	255.8	67.8	0.9
R-GMPQ	3.5	7.2	258.5	68.5	1.1
ResNet50					
Baseline	97.5	3952.6	–	76.4	–
HAQ	12.2	50.3	78.6	75.5	67.7
BP-NAS	13.4	55.2	71.7	76.7	30.2
GMPQ	12.4	53.0	74.6	76.7	2.2
R-GMPQ	10.6	51.8	76.3	76.8	2.5
HMQ	15.6	37.7	104.8	75.5	49.4
BP-NAS	11.3	33.2	119.0	75.7	35.6
GMPQ	9.6	30.7	128.6	75.8	2.7
R-GMPQ	7.9	30.1	131.5	76.2	3.1
EdMIPS	13.9	15.6	254.2	72.1	26.5
EdMIPS-C	13.7	16.0	247.2	65.6	2.9
GMPQ	8.8	15.7	252.2	73.6	3.4
R-GMPQ	10.2	15.7	251.8	74.1	3.8
MobileNet-V2					
Baseline	13.4	337.9	–	71.9	–
RQ	2.7	11.9	28.4	68.0	–
GMPQ	1.4	10.4	32.6	71.5	1.7
R-GMPQ	2.2	9.9	34.1	71.8	1.9
HAQ	1.4	8.3	41.0	69.5	51.1
HAQ-C	1.6	8.1	41.6	62.7	4.5
DJPQ	1.9	7.9	43.0	69.3	12.2
GMPQ	1.2	7.4	45.8	70.4	2.6
R-GMPQ	1.1	7.2	46.7	70.9	2.9
HMQ	1.7	5.2	64.4	70.9	33.5
DQ	1.7	4.9	68.7	69.7	21.6
SDQ	1.8	4.9	68.7	71.9	15.8
GMPQ	1.0	4.8	69.7	70.1	2.8
R-GMPQ	1.3	4.7	72.2	70.5	3.1

tency and enhance the accuracy-complexity trade-offs across different datasets.

5.3 Comparison with the Existing Mixed-precision Quantization

In this section, we compare our GMPQ with the state-of-the-art fixed-precision models containing APoT (Li et al., 2020b) and RQ (Louizos et al., 2018) and mixed-precision networks including ALQ (Qu et al., 2020), HAWQ (Dong et al., 2019c), HAQ (Wang et al., 2019a), EdMIPS (Cai and Vasconcelos, 2020), BP-NAS (Yu et al., 2020), HMQ (Habi et al., 2020), DQ (Uhlich et al., 2019) and SDQ Huang et al. (2022) on ImageNet for image classification and on PASCAL VOC and COCO for object detection. We also provide the performance of full-precision models for reference. The accuracy-complexity trade-offs of baselines are copied from their original papers or were obtained by our implementation with the officially released code, and the search cost was evaluated by re-running the compared methods. We searched the optimal quantization policy on CIFAR-10 and deployed the mixed-precision models on ImageNet, PASCAL VOC and COCO.

Table 6 The top-1 accuracy (%) on ImageNet, parameter storage cost (M), model computational cost (G) and the search cost (GPU hours) for vision transformer architectures. The training cost of Deit-T/S is 86.2 and 202.7 GPU hours.

Methods	Param.	BOPs	Comp.	Top1	Cost.
DeiT-T					
Baseline	5.11	1304.7	–	72.2	–
HAQ	0.50	20.9	62.4	62.6	27.5
EdMIPS	0.45	23.3	56.0	62.1	11.1
GMPQ	0.43	20.7	63.0	63.9	0.7
R-GMPQ	0.43	20.5	63.6	64.2	0.8
HAQ	0.63	31.7	41.2	67.5	26.4
EdMIPS	0.62	32.3	40.4	67.7	11.7
GMPQ	0.60	30.0	43.5	68.0	0.8
R-GMPQ	0.61	30.2	43.2	68.5	0.9
DeiT-S					
Baseline	22.10	4694.2	–	79.9	–
HAQ	2.07	58.6	80.1	72.5	73.2
EdMIPS	1.86	64.7	72.6	72.7	41.2
GMPQ	1.80	56.4	83.2	74.1	1.6
R-GMPQ	1.85	58.5	80.2	74.9	1.8
HAQ	2.69	93.2	50.4	75.8	72.7
EdMIPS	2.64	93.3	50.3	75.7	42.5
GMPQ	2.58	90.7	51.8	76.4	1.5
R-GMPQ	2.56	92.9	50.5	76.7	1.6

5.3.1 Image Classification

Results on ImageNet: Table 5 illustrates the comparison of storage and computational cost, the compression ratio of BOPs, the top-1 accuracy and the search cost across different architectures and mixed-precision quantization methods. HAQ-C and EdMIPS-C demonstrate that we leveraged HAQ and EdMIPS that searched the quantization policy on CIFAR-10 and directly evaluated the obtained quantization policy on ImageNet. By comparing the accuracy-complexity trade-off with the baseline methods for different architectures, we conclude that our GMPQ achieves the competitive accuracy-complexity trade-off under various resource constraint with significantly reduced search cost.

Table 7 The mAP (%) on the PASCAL VOC dataset, parameter storage cost (M), BOPs (G) and the search cost (GPU hours). The training cost for VGG16 and ResNet18 is 19.5 and 18.7 GPU hours.

Methods	Param.	BOPs	Comp.	mAP	Cost
SSD & VGG16					
Baseline	105.5	27787.7	–	72.4	–
HAQ	42.7	847.2	32.8	70.9	62.5
HAQ-C	42.9	819.7	33.9	67.6	5.1
EdMIPS	33.5	958.2	29.0	69.4	25.9
EdMIPS-C	37.2	868.4	32.0	65.2	1.5
GMPQ	36.6	796.2	34.9	70.5	1.6
R-GMPQ	32.6	761.3	36.5	70.8	1.8
HAQ	35.5	430.15	64.6	69.1	67.9
HAQ-C	32.3	445.3	62.4	66.4	6.8
EdMIPS	29.4	454.0	61.2	68.7	30.2
EdMIPS-C	31.3	423.6	65.6	64.3	1.6
GMPQ	24.7	413.5	67.2	69.2	1.8
R-GMPQ	26.9	406.8	68.3	70.3	2.0
Faster R-CNN & ResNet18					
Baseline	47.4	22534.8	–	74.5	–
HAQ	8.3	342.5	65.8	73.5	38.9
HAQ-C	8.5	337.9	66.7	70.7	4.1
EdMIPS	9.3	361.7	62.3	72.3	16.6
EdMIPS-C	8.7	348.8	64.6	69.8	0.4
GMPQ	6.4	337.9	66.7	73.9	0.5
R-GMPQ	7.2	324.7	69.4	74.3	0.6
HAQ	8.0	303.7	74.2	73.2	35.2
HAQ-C	7.6	310.4	72.6	70.4	5.2
EdMIPS	18.7	348.8	71.1	71.8	18.1
EdMIPS-C	7.4	299.3	75.3	69.2	0.4
GMPQ	6.2	286.3	78.7	73.4	0.5
R-GMPQ	6.8	284.5	79.2	73.8	0.6

Moreover, the presented R-GMPQ further enhances the trade-off with negligible extra search cost. Meanwhile, we also searched the quantization policy on CIFAR-10 directly using HAQ and EdMIPS. Although the search cost is reduced sizably, the accuracy-complexity trade-off is far from the optimal across various resource constraint, which indicates the lack of generalization ability for the quantization policy obtained by the conventional methods. Our GMPQ preserves the attribution rank consistency during the quantization policy search with acceptable computational overhead, and enables the mixed-precision quantization searched on small datasets to generalize to large-scale datasets. The presented R-GMPQ alleviates the ranking errors by smoothing attribution with hierarchical partitions and dynamically selects the optimal attribution distribution according to the sample hardness, which further enhances the generalizability of the obtained mixed-precision quantization strategy.

Our method can also be extended to other network architectures such as vision transformers. To verify this, we apply our method on DeiT-T/S (Touvron et al., 2021) to search for the optimal quantization policy. We only quantize weights and activations of fully-connected layers, and the bitwidth selection includes 2, 3 and 4 bits respectively. The results in Table 6 clearly demonstrates the superiority of GMPQ and R-GMPQ on vision transformers in mixed-precision quantization, which indicates that our method can be generalized to a wide variety of architectures with only slight modifications.

5.3.2 Object Detection

Results on PASCAL VOC: We employed the SSD detection framework with VGG16 architectures and the Faster R-CNN detector with the ResNet18 backbone to evaluate our GMPQ and R-GMPQ on object detection. Table 7 shows the results of various mixed-precision networks. Compared with the accuracy-complexity trade-off directly searched on PASCAL VOC by the state-of-the-art methods, our GMPQ and R-GMPQ acquire the competitive results with significantly reduced search cost on both detection frameworks and backbones. Meanwhile, compressing the networks with the quantization policy searched by HAQ and EdMIPS on CIFAR-10 degrades the performance significantly due to the lack of policy generalizability. Since the mixed-precision networks are required to be pretrained on ImageNet, the search cost decrease on PASCAL VOC is more sizable than that on ImageNet. Moreover, the two-stage detector Faster R-CNN has stronger discriminative power for accurate attribution generation, whose accuracy-complexity trade-off is more optimal compared with the one-stage detector due to the higher generalizability of quantization policies.

Results on COCO: Following the same detection frameworks and backbone networks on PASCAL VOC, we also evaluated our GMPQ and R-GMPQ on the COCO dataset. Table 8 depicts the accuracy-complexity trade-offs and the search cost in different computational budget. Compared with the state-of-the-art EdMIPS, GMPQ reduces the search cost by 98.5% (0.5 GPU hours vs. 32.4 GPU hours) with similar performance in the Faster R-CNN framework with ResNet-18, and R-GMPQ further enhances the policy generalizability with extra search cost of only $0.1 \sim 0.2$ GPU hours across different architectures and various model complexity. Moreover, the search cost reduction is much more sizable than that on PASCAL VOC due to the large scale of the COCO dataset.

6 Conclusion

In this paper, we have proposed a generalizable mixed-quantization method called GMPQ for efficient inference. The presented GMPQ searches the quantization

Table 8 BOPs(G), mAP@[.5, .95] and search cost (GPU hours) on COCO with state-of-the-art mixed-precision quantization methods. The average precision at different IoU thresholds and that for objects in various sizes are also illustrated. The training cost for VGG16 and ResNet18 is 56.5 and 53.2 GPU hours.

Methods	Param.	BOPs	Comp.	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	Cost
SSD & VGG16										
Baseline	105.5	27787.7	–	23.2	41.2	23.4	5.3	23.2	39.6	–
HAQ	30.6	610.7	45.5	21.4	38.9	21.0	5.5	22.4	34.0	114.8
HAQ-C	39.1	593.7	46.8	20.4	37.5	20.2	5.2	21.8	33.0	16.1
EdMIPS	21.2	621.6	44.7	20.8	38.4	20.7	5.2	22.5	33.8	38.6
EdMIPS-C	33.8	653.8	42.5	19.3	36.2	18.9	4.6	20.5	32.2	1.5
GMPQ	29.3	588.7	47.2	22.2	40.2	22.8	5.8	24.4	35.9	1.6
R-GMPQ	23.8	572.9	48.5	22.8	41.0	23.7	6.3	25.3	38.6	1.8
HAQ	12.4	445.7	62.4	20.1	37.5	19.9	5.2	21.3	32.6	95.3
HAQ-C	25.3	458.5	60.6	19.2	35.4	18.8	4.3	19.9	31.5	15.8
EdMIPS	20.5	465.5	59.7	20.8	38.4	20.7	5.2	22.5	33.8	36.7
EdMIPS-C	26.1	432.2	64.3	18.1	34.4	17.5	4.5	19.1	29.4	1.6
GMPQ	17.1	426.8	65.1	21.3	38.7	21.4	5.7	22.5	34.7	1.8
R-GMPQ	18.3	407.4	68.2	22.0	40.1	22.6	6.1	24.2	37.5	2.0
Fatser R-CNN & ResNet18										
Baseline	47.4	22534.8	–	27.6	45.7	29.1	15.3	29.2	36.2	–
HAQ	10.3	471.8	47.8	25.5	44.0	26.3	12.8	27.5	33.8	89.9
HAQ-C	11.1	529.0	42.6	22.4	38.2	23.8	11.5	24.5	29.1	9.4
EdMIPS	9.4	484.6	46.5	23.2	39.9	24.1	11.6	25.3	30.1	29.7
EdMIPS-C	9.9	508.3	44.3	22.1	39.7	22.9	11.1	23.8	30.2	0.4
GMPQ	10.3	457.1	49.3	26.8	45.7	28.1	13.8	29.3	35.0	0.5
R-GMPQ	9.1	460.8	48.9	27.1	45.9	28.6	14.6	28.6	36.0	0.6
HAQ	8.2	313.9	71.8	23.6	40.2	24.5	12.0	24.9	30.6	92.4
HAQ-C	8.4	302.9	74.4	21.6	36.9	22.6	12.1	23.2	28.7	7.6
EdMIPS	8.7	307.9	73.2	21.8	38.0	22.8	11.3	23.3	28.1	32.4
EdMIPS-C	7.5	293.8	76.7	20.4	36.0	21.0	9.5	22.0	27.5	0.4
GMPQ	7.2	285.6	78.9	25.5	44.4	26.3	12.6	27.9	33.8	0.5
R-GMPQ	6.5	290.8	77.5	26.2	44.6	27.1	14.5	27.8	35.1	0.6

policy on small datasets with attribution rank preservation, so that the acquired quantization strategy can be generalized to achieve the optimal accuracy-complexity trade-off on large-scale datasets with significant search cost reduction. We have also presented R-GMPQ that alleviates the rank errors via hierarchical attribution partitioning, and designed the dynamic capacity-aware attribution imitation to adaptively select the optimal attribution distribution for samples in various hardness. Compared with the state-of-the-art mixed-precision quantization methods, experiments have depicted that our approach achieves competitive accuracy-complexity trade-offs on image classification and object detection with significantly reduced search cost. The limitations of this framework are two-fold. First, we cannot strictly limit the BOPs of acquired quantized networks to be within the given budget as the utilized complexity loss can only minimize the model complexity without accurate control. Second, the generalizable mixed-precision quantization for other network architectures remain unexplored. Therefore, the future work contain designing proper generalization risk for supernet optimization to accurately control the model complexity, and include thoroughly verifying the effectiveness of attribution imitation on other network architectures.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0114903, and in part by the National Natural Science Foundation of China under Grant 2376032.

Data availability

Datasets used in this work are all publicly available: 1. ImageNet (Deng et al., 2009): https://www.image-net.org. 2. Pascal VOC (Everingham et al., 2010): http://host.robots.ox.ac.uk/pascal/VOC. 3. COCO (Lin et al., 2014): https://cocodataset.org. 4. CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009): https://www.cs.toronto.edu/kriz/cifar.html. 5. Cars (Krause et al., 2013): https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset. 6. Flowers (Nilsback and Zisserman, 2008): https://www.robots.ox.ac.uk/vgg/data/flowers. 7. Aircraft (Maji et al., 2013): https://www.robots.ox.ac.uk/vgg/data/fgvc-aircraft. 8. Pets (Parkhi et al., 2012): https://www.robots.ox.ac.uk/ vgg/data/pets. 9. Food (Bossard et al., 2014): https://www.kaggle.com/datasets/kmader/food41.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals.

Appendix

A. Visualization of Optimal Quantization Policy

We searched the quantization policy on different small datasets with various architectures via the presented GMPQ. Figure 11 demonstrates the optimal bitwidth allocation for weights and activations of each layer, where ResNet18 was compressed and the policy was searched on various small datasets including CIFAR-10 (Krizhevsky et al., 2009), Cars (Krause et al., 2013), Flowers (Nilsback and Zisserman, 2008), Aircraft (Maji et al., 2013), Pets (Parkhi et al., 2012) and Food (Bossard et al., 2014). Figure 12 depicts the obtained quantization strategy searched on CIFAR-10 with MobileNet-V2 (Sandler et al., 2018), ResNet18 (He et al., 2016) and ResNet50 architectures. The BOPs limit was set to 7.4G, 15.3G and 30.7G for MobileNet-V2, ResNet18 and ResNet50.

For quantization policy searched on different small datasets, the optimal bitwidth allocation varies significantly although the complexity of the obtained model is close to each other. It is observed that activations are usually assigned with higher bitwidths than weights in most quantization policy, indicating that the classification performance and attribution rank consistency are more sensitive to activation quantization than weight quantization. The bitwidth distribution of weights and activations obtained on Cars, Aircraft, Food, and CIFAR-10 is similar, which also achieves better generalization performance on largescale datasets compared with that searched on Flowers and Pets. For the Flowers and Pets datasets, the optimal quantization policy is similar to uniform quantization in fixed-precision networks, which also leads to worse accuracy-complexity trade-offs due to the lack of generalization ability.

For quantization policy for different architectures, it is observed that Layer 7, 12 and 17 in ResNet18 containing residual connections require the larger bitwidth compared with their corresponding regular branches. Since MobileNet-V2 is very compact, it receives higher bitwidths allocations than other network architectures. On the contrary, ResNet50 is compressed with lower bitwidths due to the significant redundancy compared with MobileNet-V2.

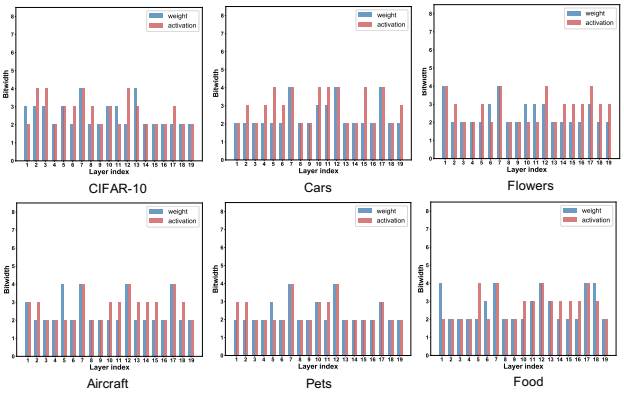


Fig. 11 The visualization of the optimal quantization policy searched on different small datasets including CIFAR-10, Cars, Flowers, Aircraft, Pets and Food.

B. Accuracy of Quantization Policy Searched on Different Small Datasets

In this section, we show the top-1 accuracy and BOPs on ImageNet of our GMPQ with the quantization policy searched on different small datasets including CIFAR-10, Cars, Flowers, Aircraft, Pets and Food. The applied network architectures contain MobileNet-V2, ResNet-18 and ResNet-50, and more accuracy-complexity trade-offs for ResNet-18 are demonstrated in Figure 10(b). Table 9 illustrates the accuracy and the complexity on ImageNet, where those of full-precision networks are also provided. The search cost is significantly reduced across various architectures compared with conventional mixed-precision quantization methods shown in Table 5, while the accuracy is only degraded slightly. The accuracy of quantization policy searched on CIFAR-10 achieves the highest, because the gap of object category between CIFAR-10 and ImageNet is the smallest compared with other datasets. Although the discrepancy of object class distribution between ImageNet and the small datasets such as Aircraft is non-negligible, the accuracy of the mixed-precision networks is still comparable with state-of-the-art approaches shown in Table 5 due to the attribution rank preservation.

C. Explanation of the Generalization Risk (9)

As visualized in Figure 3 of the manuscript, quantized networks with lower capacity tend to acquire more concentrated attribution although the attribution rank remains similar, where the networks focus on smaller regions to avoid capacity insufficiency for image representation. To further demonstrate the soundness of the observation, we report the entropy of attribution for networks in different bitwidths that reveals the attribu-

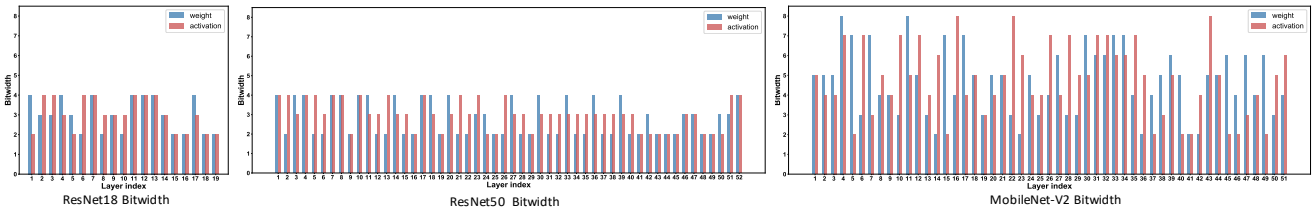


Fig. 12 The visualization of the optimal quantization policy searched on CIFAR-10 by our GMPQ. We evaluated our method with MobileNet-V2, ResNet18 and ResNet50 on ImageNet for image classification.

Table 9 Top-1 accuracy (%) and BOPs (G) on ImageNet of the mixed-precision networks searched on different small datasets across various network architectures.

Architecture	Full-precision		CIFAR-10		Cars		Flowers		Aircraft		Pets		Food	
	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs	Top1	BOPs
MobileNet-V2	71.9	337.9	70.4	7.4	69.8	7.2	67.8	7.9	69.9	7.5	66.7	7.8	69.9	7.1
ResNet18	69.7	1853.4	69.9	15.3	69.6	16.4	68.7	14.9	69.5	14.8	67.9	17.2	16.6	69.2
ResNet50	76.4	3952.6	75.8	30.7	75.5	29.8	73.8	33.2	75.6	29.5	73.3	34.1	75.6	32.7

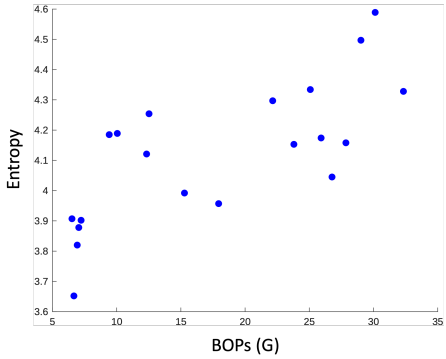


Fig. 13 The relation between the attribution entropy and the model complexity, and they are significantly positively correlated.

tion concentration. The entropy E is defined as follows:

$$E = \sum_{i,j} -M_{ij}[y_x] \log M_{ij}[y_x] \tag{15}$$

Large entropy indicates more diverse attribution and vice versa. Figure 13 shows the average attribution entropy and BOPs across the validation set of ImageNet dataset for ResNet18 in networks quantized by different optimal quantization policies (searched on ImageNet). The correlation coefficient is 0.733 between the attribution entropy and network BOPs, which verifies the observation that networks with smaller capacity acquire more concentrated attribution. For the value of p in (9), excessively large p for attribution imitation leads to over-concentrated attribution. Therefore, the networks focus on small image regions with little information, and the network capacity is not fully utilized for feature representation. On the contrary, extremely small p for attribution imitation results in attribution divergence, and focusing on large image regions causes the capacity insufficiency in the forward pass. Therefore,

Table 10 The accuracy-complexity trade-offs for different definition of p in (9).

Methods	Param.	BOPs	Comp.	Top1	Cost.
Baseline	46.8	1853.4	–	69.7	–
Squareroot	4.0	7.3	253.9	67.6	0.9
Linear	3.7	7.2	255.8	67.8	0.9
Square	3.6	7.5	247.1	67.5	0.9

we require the attribution rank of quantized and full-precision networks to be similar, while the attribution concentration is adjusted according to the network capacity. We also conducted ablation studies to show the effectiveness of the definition shown in (9). We leverage two other functions to acquire p based on the average bitwidth of the networks in the following:

$$\begin{aligned} p &= \frac{1}{L}\sum_{k=1}^{L}[Q_w^0/(\sum_{i=1}^{N_w^k}\pi_{w,i}^k q_{w,i}^k)]^{\frac{1}{2}} \cdot [Q_a^0/(\sum_{i=1}^{N_a^k}\pi_{a,i}^k q_{a,i}^k)]^{\frac{1}{2}} \\ p &= \frac{1}{L}\sum_{k=1}^{L}[Q_w^0/(\sum_{i=1}^{N_w^k}\pi_{w,i}^k q_{w,i}^k)]^{2} \cdot [Q_a^0/(\sum_{i=1}^{N_a^k}\pi_{a,i}^k q_{a,i}^k)]^{2} \end{aligned} \tag{16}$$

where the concavity is different for these functions. Table 10 shows the accuracy-complexity trade-off for ResNet18 on the validation set of ImageNet, where the linear form shown in (8) of the manuscript achieves the best performance.

D. Accuracy during the Compression Policy Search

We optimized the supernet containing all bitwidth selections with and without the generalization risk shown in (7) respectively, where we leveraged the training set from CIFAR-10 for policy search and validation set

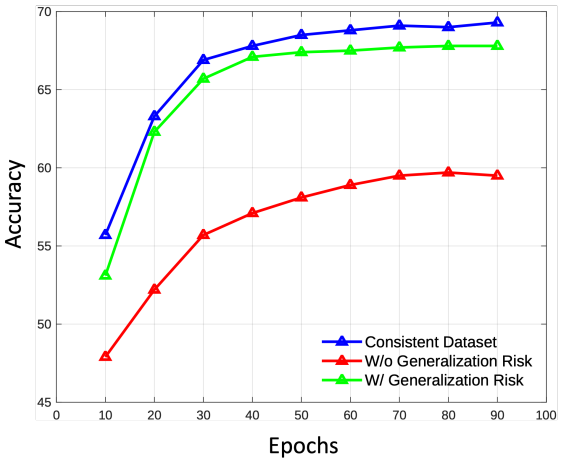


Fig. 14 The classification accuracy for quantization policies acquired during the search process.

from ImageNet for evaluation. Meanwhile, we also directly utilized the training set of ImageNet for optimizing the supernet, and report the accuracy curve for reference as the baseline. We leveraged ResNet18 as the backbone architecture, and the BOPs budget was set as 7.5G. Evaluating the acquired quantization policy requires extremely high cost because we have to finetune the quantized models until convergence. Therefore, we evaluated the searched quantization policies from different experimental settings every 10 epochs during the search process, where the quantization policy with the largest importance weight is selected for evaluation. We plot the accuracy curve in Figure 14, where our the objective with the generalization risk consistently outperforms the one without the generalization risk. The advantages of our method become more significant when the search process gradually converges. Meanwhile, the gap between our method and the optimal compression policy acquired by searching with ImageNet is small. The results can empirically verify the higher generalization ability of our method.

E. Influence of the Sample Size of Datasets for Policy Searching

In order to analysis the influence of dataset sample size for policy searching, we searched the mixed-precision quantization policies with different sample sizes on CIFAR-10. The data amount is set to be 20%, 40%, 60%, 80% and 100% of the original training set, and we report the accuracy-complexity trade-offs on the ImageNet. Moreover, we also demonstrate the performance of the optimal quantization policy that is obtained by searching on full training datasets of ImageNet. The networks quantized with acquired policies are finetuned by the datasets for evaluation. The results bring us following conclusions:

- Utilizing extremely small amount of data (e.g. $\leqslant$ 40%) from CIFAR-10 usually leads to the over-fitting for quantization policy. Since the accuracy gap between the acquired quantization policy and the optimal one is large, the quantization policy search faces the over-fitting problem.
- Enlarging the size of the dataset for quantization policy search can alleviate the over-fitting of the acquired bitwidth assignment between policy search and model deployment, since we observe that the model achieves similar accuracy for optimal quantization policy and those searched on the full training set of CIFAR-10.

F. Formulation of Rank Errors caused by Attribution Noise

The attribution acquired by Grad-cam contains noise Selvaraju et al. (2017); Sundararajan et al. (2017) which changes the attribution value slightly. However, the errors on the attribution map are significantly amplified by the ranking operation, which deviates the attribution rank of full-precision networks from the correct one obviously in attribution imitation. The generalization risk shown in (8) can be expanded as:

$$\begin{aligned}\mathcal{R}_G(\boldsymbol{W},\mathcal{Q},\boldsymbol{x}) &= \sum_{i,j}||r(M_{q,ij}[y_x]) - r(M_{f,ij}[y_x])||_2^2 \\ &= \sum_{i,j}||r(M_{q,ij}[y_x]) - r(M_{f,ij}[y_x]+\delta_{ij})||_2^2 \\ &+ ||r(M_{f,ij}[y_x]+\delta_{ij}) - r(M_{f,ij}[y_x])||_2^2 \end{aligned} \tag{17}$$

where δ_{ij} means the noise of the attribution satisfying Gaussian distribution with zero mean and σ_{ij} standard deviation. The cross term in the expansion is regarded to be zero for omission because there is no statistical correlation between the attribution value and the noise. The first term in (17) is the objective that we aim to optimize, and the second term can be represented as the KL-divergence between distribution of the two ranking variables. The KL-divergence can be written as:

$$\begin{aligned} &D_{KL}(p(r(M)=k)||p(r(M+\delta)=k)) \\ &= \int_M p(r(M)=k)\log\frac{p(r(M)=k)}{p(r(M+\delta)=k)}dM \\ &= \int_M p(r(M)=k)\cdot\frac{\partial p(r(M)=k)}{\partial M}\cdot\frac{\delta}{M}dM \\ &= C_0\delta \end{aligned} \tag{18}$$

Table 11 The accuracy-complexity trade-off across different sample size from CIFAR-10 for search.

	20%	40%	60%	80%	100%	Optimal
Accuracy	63.5	65.7	66.6	67.4	67.8	67.9
BOPs	7.5G	7.3G	7.5G	7.2G	7.2G	7.3G

Table 12 The accuracy-complexity trade-off across different numbers of top-k pixels for attribution rank preservation.

k	100	200	500	1000	2000
Accuracy	67.1	66.6	67.8	67.7	67.5
BOPs	7.3G	7.4G	7.2G	7.5G	7.4G

where we omit the subscript i and j for simplification. All variables related to M and k is deterministic when optimizing (17), and we treat them as a constant C_0. Minimizing the second term in (17) equals to minimizing the KL-divergence shown in (18), which is also equivalent to minimizing the standard deviation σ in the Gaussian distribution of δ. As semantically similar pixels usually have feature importance in similar distribution, we smooth attribution of these pixels by averaging their attribution value. Therefore, the standard deviation of their noise can be reduced since they are i.i.d. In conclusion, leveraging semantically similar pixels for attribution smoothing can reduce rank errors caused by attribution noise, which provides accurate guidance for quantized models to locate attribution correctly.

G. Ablation Study w.r.t. the Number of Pixels in Attribution Rank Preservation

Since many attribution pixels have the extremely low value (less than 10^{-2}), imitating these pixels on full-precision networks for quantized ones cannot bring sufficient supervision. The reason is that noise makes most contribution to the attribution pixel value in these cases. Therefore, we only select the top pixels based on their attribution values when minimizing the attribution distance between quantized and full-precision networks, so that the informative localization ability instead of noise in the full-precision networks is mimicked by quantized ones. In order to assign the optimal value of k for selecting top-k pixels in attribution imitation learning, we conducted ablation studies with respect to k and report the accuracy-complexity trade-offs in Table 12. Small k fails to acquire sufficient information on the full-precision attribution, and large k brings much noise in attribution imitation. Both of them degrade the trade-off between the accuracy and model complexity.

H. Details of Small Datasets for Quantization Policy Search

We introduce the datasets that we carried experiments on. For quantization policy search, we employed the small datasets including CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), Cars (Krause et al., 2013), Flowers (Nilsback and Zisserman, 2008), Aircraft (Maji et al., 2013), Pets (Parkhi et al., 2012) and Food (Bossard et al., 2014). CIFAR-10 contains $60,000$ images divided into 10 categories with equal number of samples, and CIFAR-100 contains the same number of images which are evenly distributed in 100 classes. Flowers has 8,189 images spread over 102 flower categories. Cars includes $16,185$ samples with 196 types at the level of maker, model and year, and Aircraft contains $10,200$ collected images with 100 samples for each of the 102 aircraft model variants. Pet was created with 37 dog and cat categories with 200 images for each class, and Food contains $32,135$ high-resolution food photos of menu items from the 6 restaurants.

I. Rank Errors for Different Settings for R-GMPQ

Because the full-precision attribution can be affected by noise in network training, the attribution rank may fail to reflect true region importance especially for attribution pixels with similar values. Therefore, we leverage the smoothing techniques to eliminate the noise in the attribution. Since the rank of the true attribution without noise is intractable, we randomly sampled five seeds to train the full-precision networks and used their average attribution as the approximated true attribution. We have reported the attribution rank difference on ImageNet with ResNet18 in Table 13. By comparing Table 13 with Table 2, we know that low attribution rank difference leads to better accuracy-complexity trade-offs.

J. Explanation of Attribution Similarity for Generalizable Quantization Policy Search

Let us assume that Q_D and Q_S are respectively the optimal quantization policies searched on the data in deployment and on our tractable small datasets. The

Table 13 The average rank difference between quantized attribution and approximated true attribution with different smoothing techniques and settings.

Hierarchies	RR ratio	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.5$
One-level	-	15.98	12.23	13.91	16.42
Two-level	2	17.78	13.02	14.25	17.73
	4	17.69	13.71	14.33	18.24
Three-level	2	17.93	13.57	17.12	18.59
	4	18.87	14.10	17.34	18.72
Five-level	2	19.25	18.41	19.73	20.12

generalization ability of the acquired quantization policy can be demonstrated by the difference between the expected loss of models quantized by Q_D and Q_S:

$$J = ||L(Q_D, X_{val}) - L(Q_S, X_{val})|| \tag{19}$$

where X_{val} represents the distribution of validation data in deployment. $L(Q, X)$ means the loss function of the neural networks with the quantization policy Q on the dataset X. Smaller J indicates higher generalization ability of our policy Q_S because the loss is more similar to that of the model quantized by Q_D. We expand J as follows:

$$\begin{aligned} J = &||L(Q_D, X_{val}) - L(R, X_{val}) + L(R, X_{val}) - \\ &L(R, X_{sma}) + L(R, X_{sma}) - L(Q_S, X_{sma}) + \\ &L(Q_S, X_{sma}) - L(Q_S, X_{val})|| \\ &\leqslant ||L(Q_D, X_{val}) - L(R, X_{val})|| + ||L(R, X_{sma}) - \\ &L(Q_S, X_{sma})|| + ||\big(L(R, X_{val}) - L(R, X_{sma})\big) + \\ &\big(L(Q_S, X_{sma}) - L(Q_S, X_{val})\big)|| \\ &= J_1 + J_2 + J_3 \end{aligned} \tag{20}$$

The first term J_1 is the intractable loss gap for validation data in deployment caused by quantization, and it can be regarded as a constant C_0. The second term J_2 corresponds to the loss gap for training data of our small datasets caused by quantization, and we have minimized this term in our method by optimizing the task risk. The third term J_3 can be rewritten as follows:

$$\begin{aligned} J_3 &= || \int_{X_{sma}}^{X_{val}} \frac{\partial L(R, X)}{\partial X} dX - \int_{X_{sma}}^{X_{val}} \frac{\partial L(Q_S, X)}{\partial X} dX || \\ &\leqslant \int_{X_{sma}}^{X_{val}} || \frac{\partial L(R, X)}{\partial X} - \frac{\partial L(Q_S, X)}{\partial X} || dX \end{aligned} \tag{21}$$

where $\partial L(R, X)/\partial X$ and $\partial L(Q_S, X)/\partial X$ demonstrate the attribution of full-precision and quantized models respectively. Since we require the similar attribution between quantized and full-precision models, J_3 is also minimized so that the generalization ability of the acquired quantization policy is enhanced.

References

Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, pages 2874–2883, 2016.

Joseph Bethge, Christian Bartz, Haojin Yang, Ying Chen, and Christoph Meinel. Meliusnet: Can binary neural networks achieve mobilenet-level accuracy? *arXiv preprint arXiv:2001.05936*, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *CVPR*, pages 2349–2358, 2020.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. *arXiv preprint arXiv:1404.0736*, 2014.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, pages 4312–4321, 2019a.

Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *arXiv preprint arXiv:1911.03852*, 2019b.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*, pages 293–302, 2019c.

Bo Du, Wei Xiong, Jia Wu, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Stacked convolutional denoising auto-encoders for feature representation. *IEEE transactions on cybernetics*, 47(4):1017–1027, 2016.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

Junyu Gao, Xinhong Ma, and Changsheng Xu. Learning transferable conceptual prototypes for interpretable unsupervised domain adaptation. *arXiv preprint arXiv:2310.08071*, 2023.

Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *ICCV*, pages 4852–4861, 2019.

Hai Victor Habi, Roy H Jennings, and Arnon Netzer. Hmq: Hardware friendly mixed precision quantization block for cnns. *arXiv preprint arXiv:2007.09952*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017a.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *ICCV*, pages 1389–1397, 2017b.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

Wen-bing Huang, Fu-chun Sun, et al. Building feature space of extreme learning machine with sparse denoising stacked-autoencoder. *Neurocomputing*, 174:60–71, 2016.

Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Hu Xianghong, Jeffry Wicaksana, Eric Xing, and Kwang-Ting Cheng. Sdq: Stochastic differentiable quantization with mixed precision. In *ICML*, pages 9295–9309, 2022.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS*, pages 4114–4122, 2016.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *CVPR*, pages 2810–2819, 2019.

Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group sparsity: The hinge between filter pruning and decomposition for network compression. In *CVPR*, pages 8018–8027, 2020a.

Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: A non-uniform discretization for neural networks. *ICLR*, 2020b.

Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. Runtime neural pruning. In *NIPS*, pages 2178–2188, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *ECCV*, pages 722–737, 2018.

Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. *arXiv preprint arXiv:1810.01875*, 2018.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *CVPR*, pages 11264–11272, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

Geondo Park, June Yong Yang, Sung Ju Hwang, and Eunho Yang. Attribution preservation in network compression for reliable network interpretation. *arXiv preprint arXiv:2010.15054*, 2020.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *CVPR*, pages 2250–2259, 2020.

Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In *ICCV*, pages 6718–6727, 2019.

Zhongnan Qu, Zimu Zhou, Yun Cheng, and Lothar Thiele. Adaptive loss-aware quantization for multi-bit networks. In *CVPR*, pages 7988–7997, 2020.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, pages 525–542, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint*

arXiv:1312.6034, 2013.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328, 2017.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021.

Stefan Uhlich, Lukas Mauch, Kazuki Yoshiyama, Fabien Cardinaux, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Differentiable quantization of deep neural networks. *arXiv preprint arXiv:1905.11452*, 2019.

Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. Bayesian bits: Unifying quantization and pruning. *arXiv preprint arXiv:2005.07093*, 2020.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *CVPR*, pages 8612–8620, 2019a.

Peisong Wang, Qiang Chen, Xiangyu He, and Jian Cheng. Towards accurate post-training network quantization via bit-split and stitching. In *ICML*, pages 9847–9856, 2020a.

Tianzhe Wang, Kuan Wang, Han Cai, Ji Lin, Zhijian Liu, Hanrui Wang, Yujun Lin, and Song Han. Apq: Joint search for network architecture, pruning and quantization policy. In *CVPR*, pages 2078–2087, 2020b.

Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5345–5352, 2019b.

Ying Wang, Yadong Lu, and Tijmen Blankevoort. Differentiable joint pruning and quantization for hardware efficiency. In *ECCV*, pages 259–277, 2020c.

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *ICCV*, pages 7639–7648, 2021.

Ziwei Wang, Jiwen Lu, Chenxin Tao, Jie Zhou, and Qi Tian. Learning channel-wise interactions for binary convolutional neural networks. In *CVPR*, pages 568–577, 2019c.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *CVPR*, pages 1161–1170, 2020.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, pages 501–509, 2019.

Haichuan Yang, Shupeng Gui, Yuhao Zhu, and Ji Liu. Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach. In *CVPR*, pages 2178–2188, 2020.

Haibao Yu, Qi Han, Jianbo Li, Jianping Shi, Guangliang Cheng, and Bin Fan. Search what you want: Barrier panelty nas for mixed precision quantization. In *ECCV*, pages 1–16, 2020.

Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *CVPR*, pages 7370–7379, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *IJCV*, 126(10):1084–1102, 2018.

Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *ICML*, pages 7543–7552, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. In *CVPR*, pages 3233–3242, 2021.